UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

PREDICCIÓ DE COTITZACIÓ D'ACCIONS MITJANÇANT GRAPH NEURAL NETWORKS(GNN): BENCHMARKING DE GRAFS RELACIONALS D'EMPRESES

MIQUEL MUÑOZ GARCÍA-RAMOS

Director/a
SERGI ABADAL CAVALLÉ (Departament d'Arquitectura de Computadors)

Codirector/a
AXEL TOMAS WASSINGTON (Departament d'Arquitectura de Computadors)

Titulació
Grau en Enginyeria Informàtica (Computació)

Memòria del treball de fi de grau

Facultat d'Informàtica de Barcelona (FIB)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

Índex

Llista de taules

Llista de figures

Resum

En aquest de treball de fi de grau es desenvolupa un *benchmark* per a l'avaluació de grafs relacionals d'empreses del *SP500* per al problema de predicció d'accions borsàries. Múltiples mètodes de construcció de grafs relacionals són avaluats amb fonts de la indústria, notícies, estructura, inversors institucionals, correlacions d'indicadors, etc. Per desenvolupar el *benchmark*, s'ha reconstruït una *Graph Neural Network* temporal heterogènia [16], que és capaç de batre al mercat consistentment amb el pas del temps. A més, s'ha proposat un mètode d'avaluació del graf tenint en compte el graf de correlacions del període següent. D'aquesta manera, es mesura la similitud entre grafs i pot facilitar la recerca de mètodes de construcció de grafs relacionals en futures investigacions.

Resumen

En este de trabajo de fin de grado se desarrolla un *benchmark* para la evaluación de grafos relacionales de empresas del *SP500* para el problema de predicción de acciones bursátiles. Múltiples métodos de construcción de grafos relacionales son evaluados con fuentes de la industria, noticias, estructura, inversores institucionales, correlaciones de indicadores, etc. Para desarrollar el *benchmark*, se ha reconstruido una *Graph Neural Network* temporal heterogénea [16], que es capaz de batir al mercado consistentemente con el paso del tiempo. Además, se ha propuesto un método de evaluación del grafo teniendo en cuenta el grafo de correlaciones del periodo siguiente. De este modo, se mide la similitud entre grafos y puede facilitar la investigación de métodos de construcción de grafos relacionales en futuras investigaciones.

Abstract

In this thesis, a benchmark is developed for the evaluation of relational graphs of SP500 companies for the stock market prediction problem. Multiple methods of constructing relational graphs are evaluated with industry sources, news, structure, institutional investors, correlations of indicators, etc. To develop the benchmark, a heterogeneous temporal Graph Neural Network has been reconstructed [16], which is able to beat the market consistently over time. Furthermore, a method of evaluating the graph by taking into account the correlation graph of the following period has been proposed. In this way, the similarity

between graphs is measured and can facilitate the investigation of relational graph construction methods in future research.

Agraïments

Vull agrair al Sergi Abadal (director del TFG) i l'Axel Wassington (codirector del TFG) pel seu suport. Des del moment previ a triar el tema del TFG, aquesta alternativa es presentava com la més arriscada en comparació a altres projectes iniciats al *Barcelona Neural Networking Center* (BNN) d'aplicacions de GNNs. Tot i això, es van disposar a dirigir un treball aplicat a finances, que es troba fora del seu àmbit d'investigació, i van dipositar la confiança necessària per desenvolupar un treball d'aquest estil. Sobretot, en l'etapa on es va investigar l'estat de l'art i es va intentar reconstruir diversos models sense èxit, la seva orientació va ser fonamental per a conduir el treball.

Per altra banda, les xerrades amb els amics i companys de carrera, Marc Díaz i Tomás Osarte, al bar de la FIB sobre els avanços dels TFGs, van enriquir de nous punts de vista i observacions del treball.

1 Introducció

1.1 Contextualització

En el moment en el qual ens trobem, on els mètodes basats en la col·lecció i interpretació de dades creixen acceleradament, un dels àmbits en què ha despertat més interès i popularitat, és en el camp de l'aprenentatge automàtic. Més en concret, en els models que tenen la capacitat de modelar dades com a grafs. Aquestes dades poden ser provinents de tota mena d'àrees del coneixement com la física, les ciències socials, la logística de transportació o les finances.

Un sistema financer pot ser definit com un sistema complex amb molts components que tenen relacions sofisticades, el qual s'adapta i reflecteix els esdeveniments que succeeixen a la realitat. Per aquestes raons, podem pensar en un mercat financer com en un graf extremadament complex, on existeixen incentius lucratius en el desenvolupament de models predictius que aconsegueixin anticipar el seu comportament.

Mitjançant la tasca de classificació de nodes, el desenvolupament de models predictius amb *Graph Neural Network* (GNN), pot ser emprat per a tota classe d'àmbits i les finances no són una excepció. Alguns exemples són el risc d'incompliment de préstecs, la detecció de frau financer [1] o en el problema que es basa aquest treball de fi de grau: la predicció d'accions en els mercats borsaris. Cal dir, que les xarxes neuronals tradicionals com les *Graph Convolutional Networks* (GCN), *Recurrent Neural Networks* (RNN) o *Long short-term memory* (LSTM) no són capaces d'aprendre d'informació que té una relació arbitrària en forma de graf, com són les finances. Per la qual cosa, les GNNs tot i que són relativament noves, són aplicables i comunament utilitzades per a aquest problema. No obstant això, donada l'alta complexitat i competitivitat en els mercats financers, el desenvolupament d'aquests algoritmes per a la identificació de relacions no lineals, comporta diversos reptes en el processament de *features* temporals i de relacions heterogènies entre accions.

Una de les fases més significatives en el rendiment de les GNNs és la construcció del graf. Per aquesta raó, aquest treball pretén investigar l'impacte de diferents mètodes de construcció i utilitzar diferents mètriques d'avaluació del graf per a la creació d'un *benchmark*. Aquest treball de fi de grau (TFG), se situa en el marc de la *Facultat d'Informàtica de Barcelona* (FIB) de la *Universitat Politècnica de Catalunya* (UPC) i més en concret, en el grup d'investigació *Barcelona Neural Networking Center* (BNN).

1.2 Història de les finances quantitatives

Aquest punt s'ha considerat especialment rellevant, atès que existeix molt misticisme i escepticisme en aquest camp. Fins i tot, en professionals i docents de l'àmbit de l'economia o de les ciències computacionals.

L'anàlisi quantitativa és l'ús de models matemàtics i estadístics aplicats a finances i maneig d'inversions. Els treballadors d'aquesta indústria són comunament coneguts com a *quants*.

L'inici d'aquesta indústria se situa el 1900 amb la tesi doctoral de Louis Bachelier, *Theory of Speculation*. Aquesta, tractava de modelar el preu de les opcions financeres a partir d'una distribució normal. Més tard, durant el 1950 es van desenvolupar diversos avanços en adaptar conceptes matemàtics per a finances en teoria de diversificació de *portfolio* i maneig i quantificació de risc [23].

El terme *quantitative investment management* va ser primerament introduït per Edward Thorp (professor de matemàtiques a New Mexico State University (1961–1965) i University of California, Irvine (1965–1977)). Se'l considera "el pare de les finances quantitatives". Thorp va desenvolupar un sistema comunament conegut com el "comptatge de cartes" aplicat al *blackjack* (el conegut joc de cartes que solia jugar a Las Vegas). Aquest sistema basat en teoria probabilística i anàlisi estadística, el va utilitzar per crear el primer fons d'inversió d'aquest tipus, amb un èxit immediat [24].

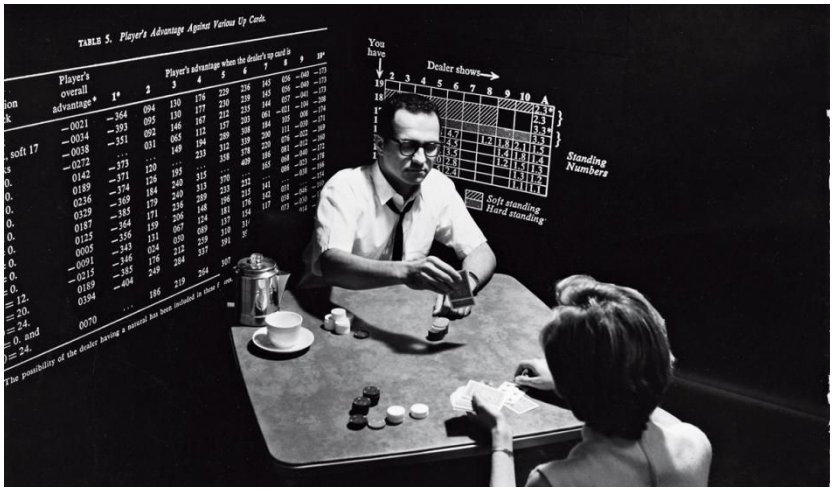


Figura 1: Ed Thorp al 1964. Font: *A Man for All Markets*, Thorp 2017. [22].

Per altra banda, a finals dels anys seixanta, Wells Fargo (una companyia estatunidenca de serveis financers) va començar a contractar més matemàtics i físics en lloc d'economistes o llicenciats en finances. Entre ells, hi havia Fischer Black. Doctorat en matemàtiques aplicades per Harvard, va sortir de Wall Street per tornar al camp acadèmic durant els anys setanta a la Universitat de Chicago. Allà, va desenvolupar el famós model *Black-Scholes* per a derivats financers, pel qual va obtenir el Premi Nobel d'Economia el 1997.

Durant els anys 70 i 80, ja estava estandarditzat a Wall Street l'ús de models quantitatius, però no es disposava de tecnologia necessària per a l'escalabilitat. El 1990, va haver-hi un increment de potència computacional exponencial que va permetre l'aparició de World Wide Web, Google i Amazon van ser fundades, la propietat d'ordinadors va créixer a més del 60% als Estats Units i els fons d'inversió van tenir un *boom* gegantí de popularitat.

No obstant això, el 1998 va col·lapsar *Long-Term Capital Management (LCTM)* a causa de la manca de pagament del deute rus, que va comportar un rescat de la Reserva Federal. Així i tot, els inversors institucionals van continuar apostant per les estratègies quantitatives. Si més no, amb el 6 d'agost de 2007 molts fons quantitatius van haver de vendre accions per cobrir pèrdues d'accions no liquidables, provocant un efecte dominó que va comportar una caiguda del mercat. Encara que els preus es van estabilitzar en una setmana, l'esdeveniment conegut com el *Quant Quake* comportar una reducció en la confiança en aquest sector. Els actius de *Goldman Sachs*, que havien arribat al pic de 165.000 milions de dòlars van arribar a caure fins a 38.000 milions el 2012.

El 2011, el *Quant Quake* va provocar el tancament del fons *Global Alpha* de *Goldman Sachs* (un fons purament quantitatiu). En canvi, l'estratègia híbrida d'humans i màquines de *Goldman Sachs* anomenada *Global Equity Insights with Country Tilts* va tenir molt millors resultats, tant el 2007 com a la crisi de la COVID-19. El mateix es pot dir per als altres fons "d'humans i màquines" que, no com els seus competidors purament quantitatius, van resistir el pas del temps. Des d'aleshores, els gestors quantitatius han reavaluat les estratègies, focalitzant-se en la transparència, la liquiditat i la importància de la supervisió d'humans, especialment, durant períodes de turbulències financeres [24]. Actualment, la cerca de mètodes quantitatius més acurats continua vigent, amb eines com les GNNs explorades en aquesta tesi. La Figura 2 mostra la línia històrica de les finances quantitatives.

Figura 2: Línia històrica de les finances quantitatives. Font: Bank Of Montreal, Canadà (2024). [24].

1.3 Definició de conceptes

A continuació, són definits els conceptes fonamentals per entendre el plantejament del projecte.

1.3.1 Intel·ligència artificial i aprenentatge automàtic

La intel·ligència artificial és la disciplina en el camp de les ciències de la computació que tracta de desenvolupar algoritmes i màquines que imitin la intel·ligència humana [2].

Un dels subcamps de la intel·ligència artificial és l'aprenentatge automàtic (*Machine Learning*). Aquesta branca, tracta de desenvolupar d'algoritmes estadístics que aprenguin a partir de dades i siguin capaços d'inferir i generalitzar dades no experimentades mitjançant el reconeixement de patrons.

Una de les alternatives emprades en l'aprenentatge automàtic és l'ús de les *Graph Neural Networks* (GNN). La branca basada en les xarxes neuronals profundes s'anomena aprenentatge profund (*Deep Learning*).

1.3.2 *Graph Neural Network* (GNN)

Les xarxes neuronals són emprades en l'àmbit de l'aprenentatge automàtic per a desenvolupar models que tinguin la capacitat de predir resultats amb un con-

trol adaptatiu. Mitjançant un conjunt de dades, els models aprenen a partir de l'experiència i infereixen conclusions a partir d'informació prèvia que pot semblar relacionada o no.

Aquestes xarxes neuronals estan inspirades en l'organització biològica de les neurones dels animals. L'homòleg de la neurona seria cada un dels nodes de la xarxa i les possibles connexions sinàptiques esdevindrien les arestes del graf.

Una *Graph Neural Network* (GNN) és un tipus de xarxa neuronal per processar dades que es poden representar com a graf.

Les GNN consten de múltiples capes, on cada una és responsable de generar una representació latent o intermèdia del node. L'intercanvi d'informació entre els nodes durant la fase d'entrenament és una de les idees fonamentals de les GNN. En cada capa es desenvolupen les següents fases:

- Transferència del missatge: cada node agrega informació provinent dels nodes veïns. El missatge estarà condicionat per les mateixes característiques del node que l'envia i dels seus nodes veïns.
- Actualització del node: cada node aprofita la informació rebuda per actualitzar la seva representació (*embedding*) a partir dels missatges rebuts dels nodes veïns.

A la Figura 3 es pot observar el procediment d'ús de GNNs per al problema de predicció d'accions borsàries.

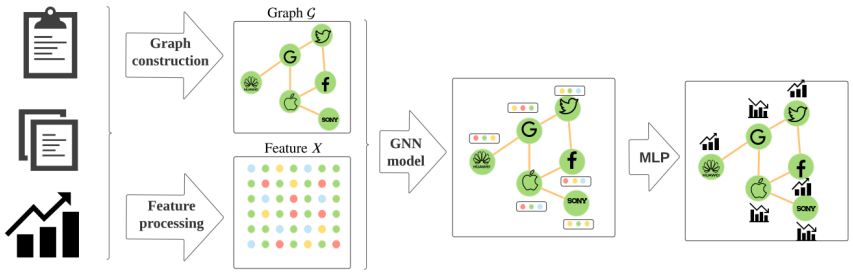


Figura 3: Desenvolupament d'una GNN per a la predicció de cotitzacions borsàries. Font: Wang, J., Zhang, S., Xiao, Y., & Song, R. (2022). *A Review on Graph Neural Network Methods in Financial Applications* [1].

1.3.3 Oferta i demanda

Tenint en compte que aquest treball de fi de grau ha estat desenvolupat a la facultat d'informàtica, s'ha considerat necessari explicar alguns principis bàsics d'economia i finances molt resumidament per poder entendre el treball amb claredat.

El famós principi d'oferta i demanda postula que, *ceteris paribus* (mantenint tot cosa altra constant), en un mercat competitiu, variarà fins que la quantitat demanda del bé iguali la quantitat oferta, arribant a l'equilibri econòmic de preu i quantitat [44].

Per una banda, l'oferta es representa gràficament com una corba que relaciona el preu del bé amb la quantitat oferta pels venedors del be en qüestió. Sota l'assumpció de perfecta competició l'oferta es determina segons el cost marginal. Un increment en el cost dels materials disminuiria l'oferta, desplaçant la corba cap a l'esquerra. La corba d'oferta té pendent positiu, ja que si el preu del producte augmenta, el venedor estaria més incentivat a produir-lo.

En canvi, la corba de demanda representa la quantitat d'un bé determinat tal que els compradors estaran disposats a comprar en diferents preus. Els consumidors compararan una unitat sempre que el valor marginal de comprar una unitat extra sigui major al preu a pagar. La corba de demanda té pendent negatiu, donat que si el preu del bé disminueix, els consumidors estaran més incentivats a comprar-lo.

Quan hi ha un increment de demanda per a un preu determinat en qualsevol mercat de béns, la corba de demanda es desplaça cap a la dreta arribant a un nou equilibri. En aquest cas, el preu i la quantitat seran més alts. Un exemple d'aquest cas en el mercat borsari, podria ser causat per uns beneficis superiors als esperats en una conferència de guanys. En cas contrari, una disminució de la demanda, comportaria una disminució tant de preu com de quantitat.

Per altra banda, si hi ha un increment d'oferta, la corba d'oferta es desplaçarà cap a la dreta, incrementant quantitat però, disminuint el preu. Un exemple d'un increment d'oferta podria ser la decisió d'una empresa de fer un *stock split*. Aquesta és la decisió d'incrementar el nombre d'accions en circulació d'una empresa mitjançant l'emissió de noves accions als accionistes existents, en una proporció establida. En cas contrari, si l'oferta disminueix, l'efecte serà contrari. La quantitat disminuiria i el preu augmentaria.

A la Figura 4 es poden identificar les corbes d'oferta i demanda. Concretament, hi ha un augment de demanda que comporta un nou equilibri (t1) amb

un increment de preu i de quantitat.

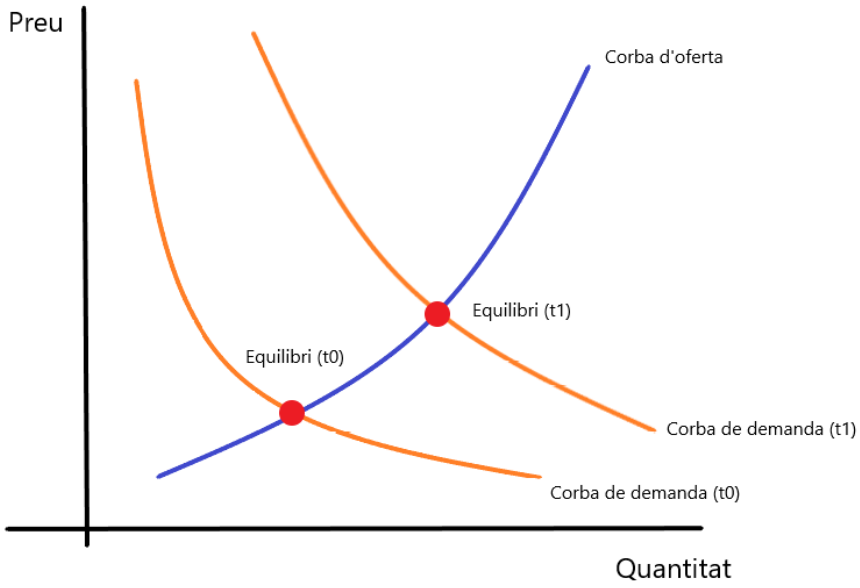


Figura 4: Augment de demanda. Font: elaboració pròpia.

1.3.4 Mercats de valors

Un mercat de valors, mercat borsari, mercat de renda variable o mercat d'accions és l'agregació de compradors i venedors d'accions. Una acció representa el dret de propietat parcial sobre les empreses. Poden incloure valors cotitzats en una borsa pública, així com accions que solament es negocien en privat, com per exemple mitjançant operacions de *venture capital* (operacions en les quals s'aporta capital a *startups* amb potencial creixement però alts nivells de risc).

La borsa és on els *stockbrokers* i inversors poden comprar i vendre accions, bons i altres valors. Moltes empreses grans cotitzen en borsa, per a una major liquiditat i, per tant, una major atracció a molts inversors. Algunes empreses cotitzen en més d'una borsa en diferents països per també, atraure inversors internacionals.

Un intercanvi en aquest mercat es significa la transferència d'una acció d'un venedor a un comprador a canvi de diners. Això requereix que ambdues parts

estiguin d'acord en el preu. Un comprador potencial ofereix un preu específic per a una acció determinada i el venedor demana un altre preu. Quan aquests preus coincideixen, es produeix una operació de venda.

El propòsit d'una borsa de valors és facilitar aquest intercanvi per constituir un mercat. Les borses proporcionen informació de negociació a temps real dels valors cotitzats per facilitar la determinació de preus.

Els participants d'un mercat poden ser tant inversors individuals com inversors institucionals (per exemple un fons d'inversió). Algunes borses són llocs físics, però també n'hi ha de virtuals com per exemple, el *NASDAQ* (borsa americana amb seu a Nova York). A la Figura 5 es pot observar la borsa de Nova York.

Figura 5: Fotografia de la borsa de Nova York. Font: Michael Nagle, *Bloomberg* [46].

1.4 Identificació del problema i estat de l'art

Tal com s'ha esmentat anteriorment al punt 1.1, existeixen múltiples aplicacions de les GNNs a diversos problemes en el camp financer. Tanmateix, ja que els conjunts de dades utilitzats per a desenvolupar models enfocats al problema de predicció de cotització d'accions són parcialment gratuïts (vegeu punt B.1), el

treball se centrarà únicament en aquest problema.

Tot i que encara està viu el debat de fins a quin punt la cotització de les accions és predictible, és un problema que compta amb una gran atenció i hi ha bastant literatura en la predicció dels moviments dels mercats utilitzant models d'aprenentatge automàtic [1, 6–11, 16, 17]. Tradicionalment, es poden distingir dues maneres d'abastar el problema: l'anàlisi tècnica i l'anàlisi fonamental. L'anàlisi fonamental utilitza informació no numèrica com notícies o conferències de guanys. Per altra banda, l'anàlisi tècnica utilitza dades numèriques com cotitzacions de tancament, cotitzacions d'obertura o volum d'inversió. La limitació d'alguns d'aquests models és que sovint tenen l'assumpció que les accions són independents. De manera que, per reconèixer aquestes dependències, cada cop més s'usa la representació de les accions com a graf on cada acció és caracteritzada com a node i una aresta existeix si hi ha una relació entre les accions. Aleshores, predir el moviment d'accions pot ser formulat com un problema de classificació de nodes on els models basats en GNNs poden ser emprats per a tractar de fer prediccions borsàries.

No obstant això, a diferència de molts altres camps on existeixen *benchmarks* de grafs, no hi ha cap manera estandarditzada de construir el graf que representi la relació entre accions [1]. Amb l'abundància de relacions existents als mercats financers, esdevé una tasca molt complicada com elaborar aquesta relació entre accions. Encara més, tenint en compte la volatilitat intrínseca dels mercats, el modelatge de *features* per a poder identificar relacions no lineals és també un repte complex. Les fonts de dades que es poden utilitzar són molt diverses: relacions de cadena de subministrament, conferències de guanys, notícies, cotitzacions, indicadors tècnics, etc. Malgrat això, no existeix un "graf ideal" perquè escauen mètodes d'avaluació de grafs i consegüentment, també manquen *benchmarks* exhaustius dels diferents mètodes de construcció de grafs. Inclús s'esmenta a l'article més punter que recull l'estat de l'art de les GNN en finances que "en el futur es podria dissenyar un mètode d'avaluació de grafs que ajudi els investigadors a construir un millor graf relacional" [1].

En conseqüència, el treball de fi de grau avalua diferents mètriques per a la construcció de grafs amb l'objectiu de la creació d'un *benchmark*. Mitjançant la construcció d'un model punter de GNN heterogènia diversos mètodes de construcció de grafs són avaluats en comparació amb un graf de referència, tant en mètriques de testatge predictives que depenen del model construït com també, en mètriques que avaluen els grafs amb independència del model utilitzat a partir del graf ideal.

És convenient entendre que el TFG en cap cas, tracta de ser un TFG orientat a identificar causalitats, donat que aquests aspectes són tractats en el camp de l'econometria i queden fora de l'abast del projecte.

A continuació s'esmenta un resum de la literatura existent en la predicció de cotització d'accions mitjançant GNNs que serviran per contrastar els resultats que s'obtinguin en el treball. Sobretot, la part més rellevant seran els diferents mètodes de construcció de grafs que es poden identificar a la Taula 1. Atès que, bona part d'aquestes estratègies de construcció seran construïdes o donaran peu a desenvolupar-ne de noves.

Considerant l'anàlisi de la construcció del graf dels diferents articles de la Taula 1, es poden extreure les següents observacions.

- L'article *Exploring Graph Neural Networks for Stock Market Predictions with Rolling Window Analysis* [6], considera 7 tipus de relacions entre empreses mitjançant *Wikidata*. S'analitza el rendiment, per separat d'aquestes relacions, però no es comparen amb altres mètodes de construcció. Es conclou, que els grafs relacionals que incorporen relacions de consumidor, proveïdor, inversor, etc. són efectius per al problema de predicció d'accions. S'aconsegueix un 29.95% de retorn de la inversió mitjà anual en 20 anys de dades per a les empreses del mercat d'accions japonès (índex *Nikkei 225*).

- L'article *Time-aware Graph Relational Attention Network for Stock Recommendation* [7] construeix el graf a partir de relacions d'indústries i sectors. No hi ha una anàlisi d'alternatives al mètode de construcció del graf utilitzat. Conclouen que el mètode utilitzat per al *ranking* d'empreses, permet predir quines obtindran un major retorn de la inversió. Obtenen entre el 2013 i el 2017 un retorn mitjà de la inversió anual del 92% per al *NASDAQ* i un 138% per al *NYSE* (principals borses dels Estats Units).

- L'article *Modeling the Stock Relation with Graph Network for Overnight Stock Movement Prediction* [8] utilitza les correlacions històriques d'indicadors del mercat entre accions per a la construcció del graf. No s'analitzen diferents tipus de construcció de grafs. Conclouen que el seu mètode de predir les accions durant la nit (mentre la borsa és tancada) a partir de notícies és un mètode efectiu. Aconsegueixen una precisió mitjana (per a la predicció de pujada o baixada de la cotització) del 56.14% i 58.71% per als índexs borsaris japonesos TPX500 i TPX100 respectivament. Utilitzen dades

d'entre el 2013 i el 2018.

- L'article *Deep Attentive Learning for Stock Movement Prediction From Social Media Text and Company Correlations* [9] utilitza relacions de *Wikidata* per a la construcció del graf. Tot i que es comparen els resultats amb l'estat de l'art, no s'aïlla la construcció del graf per a poder ser comparada amb altres mètodes de construcció. Obtenen un *sharpe ratio* (mètrica que ajusta el retorn de la inversió al risc) d'1,05. Els resultats s'obtenen testejant l'any 2015 per a les empreses del *SP500* (500 majors empreses dels Estats Units).

- L'article *FinSense: An Assistant System for Financial Journalists and Investors* [11] utilitza relacions de coocurrència de les empreses en notícies. No es compara aquest mètode de construcció del graf amb altres alternatives. Aconsegueixen una precisió mitjana de 56.77% (per a predir la pujada o baixada) en empreses de la borsa de Taiwan.

- L'article *Temporal Relational Ranking for Stock Prediction* [35] utilitza relacions d'indústria i sector per a la construcció del graf. També s'utilitzen relacions de *Wikidata*. S'analitza el rendiment d'algunes d'aquestes relacions per veure quines indústries o relacions concretes provinents de *Wikidata* són més eficients. Tot i això, no es comparen aquests mètodes de construcció del graf amb altres alternatives. Es testegen diverses estratègies d'inversió arribant a resultats de fins al 119% de retorn de la inversió a l'any 2017.

- L'article *Temporal and Heterogeneous Graph Neural Network for Financial Time Series Prediction* [16] utilitza correlacions històriques d'indicadors per a la construcció del graf. No es compara aquest mètode de construcció amb cap altre. S'obté un retorn anual de la inversió del 66.5% al *SP500* i d'un 63,2% per al CSI300 (índex borsari xinès) l'any 2020. L'arquitectura d'aquest article ha estat reconstruïda per a dur a terme el TFG (vegeu el punt 4.2).

Característiques	**Graf**	**Mètode**	**Mètrica d'avaluació**
Mitjana mòbil dels últims 5/10/20/30 dies dels preus de tancament per a les empreses llistades a *Nikkei 225*	Graf de relacions de proveïdors, clients, socis i accionistes	*Relational Stock Ranking* [6]	*Return ratio*, *sharpe ratio*
Mitjana mòbil dels últims 5/10/20/30 dies dels preus de tancament i documents de descripció d'accions per a les empreses llistades al *SP500* i *NYSE*	Xarxa de relacions basada en empreses de la Wiki	T*ime-aware graph Relational Attention Network* [7]	*Mean Squared Error*, *Mean Reciprocal Rank i Investment Return Ratio*
Notícies de les empreses llistades al *Tokyo Stock Price Index 500/100*	Graf de correlació d'accions	*LSTM Relational Graph Convolutional Network* [8]	Precisió
Informació de preus i xarxes socials per a les empreses llistades a l'índex *SP500* o als mercats *NYSE* o *NASDAQ*	Xarxa de relacions basada en empreses de la Wiki	*Multipronged Attention Network for Stock Forecasting* [9]	Precisió, F1, *Matthew's Correlation Coefficient*
Característiques de text i àudio de les *earnings call* per a les empreses a l'índex *SP500*	Graf d'*earning calls* d'accions	*Volatility forecasting via Text-Audio fusion with Graph convolution networks for Earnings calls* [10]	*Mean Squared Error, R-squared*
Notícies i atributs per a les etiquetes d'accions	Graf de co-ocurrència de notícies	*Hybrid Attention Network* [11]	Precisió, *Matthew's Correlation Coefficient*

Característiques	**Graf**	**Mètode**	**Mètrica d'avaluació**
Cotitzacions històriques d'obertura (20 dies) dels índexs *SP500* i *CSI300*	Graf de correlació d'accions	*Graf de correlació d'accions heterogeni* [16]	Precisió, retorn anual, risc, *sharpe ratio*.
Cotitzacions històriques, esdeveniments i notícies d'accions tipus "A" de la Xina	Graf heterogeni multimodal	*Multi-modality graph neural network* [17]	F1, *sharpe ratio*, retorn anual
Mitjana mòbil dels últims 5/10/20/30 per a les empreses del *NASDAQ* i *NYSE*	Xarxa de relacions Wiki i indústria-sector	*Relational Stock Ranking* [35]	*Mean Squared Error*, *sharpe ratio*, retorn de la inversió, *Mean Reciprocal Rank*

Taula 1: Taula resum de la literatura existent en la predicció de cotització d'accions mitjançant GNNs [1]. Elaboració pròpia.

1.5 Agents implicats

Aquest projecte pot interessar als investigadors i les empreses que tracten d'estudiar l'aplicabilitat de l'aprenentatge automàtic i les GNN. Especialment, a aquells que ho investiguen a l'àmbit financer donat que el treball proporciona mètodes d'avaluació i eines que els pot ajudar directament a la millora en la construcció i l'avaluació dels seus grafs.

Altrament, la predicció de les cotitzacions borsàries pot interessar no només a empreses (com fons d'inversió o *trading firms*) i institucions financeres sinó, que qualsevol individual es podria beneficiar de lucrar-se amb aquests mètodes mitjançant la inversió algorítmica.

Cal esmentar que aquest treball també suposaria un avanç en el grup de recerca del *Barcelona Neural Networking Center* (BNN), ja que s'investiga les capacitats de l'aplicabilitat de les GNN. Les persones directament implicades en el desenvolupament del projecte són:

- el Sergi Abadal (director del TFG)
- l'Axel Wassington (codirector del TFG)
- el Miquel Muñoz (autor del TFG)

2 Justificació

Encara que algunes d'aquestes raons ja s'han anat comentant anteriorment, en aquest punt es llisten de forma clara i concisa. Les raons per les quals s'ha abastat aquest problema són les següents:

- **Inexistència de *benchmarks***: no existeixen *benchmarks* per als grafs utilitzats per a modelar les relacions entre accions borsàries [1].
- **Inexistència de mètodes d'avaluació**: per conseqüència de la inexistència de *benchmarks*, no hi ha mètodes per a l'avaluació del graf independents del model [1].
- **Article punter**: la identificació del problema està clarament indicada per un article de referència punter que incita a cobrir aquest nínxol de coneixement [1].
- **Eficiència de les GNN**: s'ha demostrat que les GNN poden aportar bones prediccions en el processament i anàlisi de dades financeres, donada la capacitat de capturar relacions no lineals i patrons ocults en conjunt de dades massius [1].
- **Flexibilitat de les GNN**: les GNN poden ser adaptades i entrenades amb nous conjunts de dades per a abordar diferents tipus de problemes no experimentats fins al moment. Això inclou a dades de noves entitats o mercats financers. També hi ha la capacitat de diferents tipus de construcció dels grafs per abordar aquests problemes.
- **Millora de la precisió**: l'avanç en el coneixement i avaluació de la construcció del graf pot suposar un avanç en el rendiment de models existents i futurs. És molt significatiu, en un entorn tan dinàmic i competitiu com són els mercats financers. En concret, tot i que la literatura revisada consta de diferents mètriques d'avaluació, la precisió dels models és d'entre el 50% i el 60% [8] [9] [11] (formulant el problema com una classificació binària de pujada o baixada de la cotització). Aquests resultats superen a la predicció aleatòria que té una precisió del 50%.
- **Evolució del camp**: amb el creixent interès i avanç continu que tenen les xarxes neuronals i l'aprenentatge profund suposa una oportunitat rellevant amb un nínxol de coneixement en el qual queda molt per explorar.

Cal entendre que l'evolució de l'aprenentatge automàtic i en conseqüència l'entrada de nous algoritmes i estratègies d'inversió, afecta els algoritmes existents dels altres competidors que interactuen en el mercat. D'aquesta manera algoritmes guanyadors en el passat poden quedar obsolets amb el pas del temps. Això proporciona una oportunitat contínua per tractar de batre el mercat.

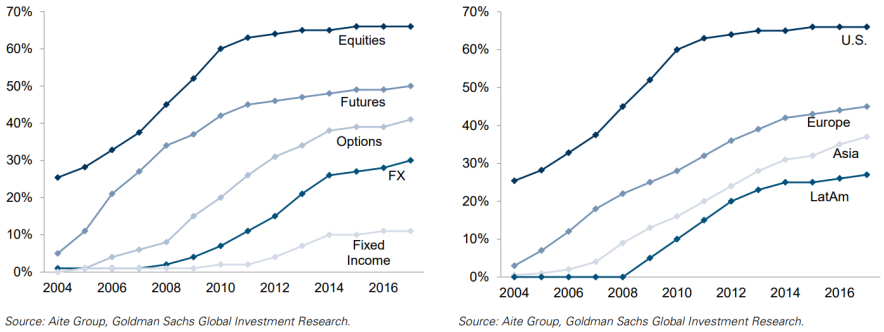


Figura 6: Percentatge d'inversió algorítmica per tipus d'actiu o de país. Font: Goldman Sachs (2017). *Top of Mind, Global Macro Research* [12].

La Figura 6, és un gràfic elaborat pel grup de banca d'inversió *Goldman Sachs* on es mostra com la inversió algorítmica està en una tendència creixent en els últims anys, arribant aproximadament fins al 70% del volum de mercat global d'accions. Aquestes dades corroboren l'auge que han tingut les ciències computacionals en aquest sector.

3 Abast

3.1 Objectius i subobjectius

Els objectius del projecte són:

- Proposar un *benchmark* per a la construcció de grafs relacionals d'accions borsàries.
- Proposar mètodes d'avaluació del graf en relació amb el *benchmark* desenvolupat. De manera que, es pugui mesurar la distància entre el graf que es vulgui construir i el "graf ideal" (graf de referència que aconsegueix un major rendiment). Així doncs, no és necessari testejar el model amb ambdós grafs ni hi ha una dependència del testatge del graf amb el model utilitzat.
- Desenvolupar o reconstruir un programari (*pipeline*) existent i punter basat en GNN per a la predicció de cotitzacions d'accions borsàries. D'aquesta manera es pot corroborar el desenvolupament del *benchmark* i de les mètriques d'avaluació entre grafs (independents del model).
- Testejar el programari amb els diferents tipus de grafs amb dades del *SP500* (índex borsari de les 500 empreses amb més capitalització de mercat dels Estats Units).

Com a subobjectius es poden destacar:

- Investigar i seleccionar diferents mètodes per a la construcció de grafs per al desenvolupament del *benchmark*: correlació d'indicadors tècnics, utilitzar les relacions de Wikidata, relacions entre inversors institucionals, notícies, etc.
- Investigar l'estat de l'art actual en el camp de l'aprenentatge automàtic aplicat a la predicció de cotitzacions d'accions borsàries.
- Desenvolupament de *crawlers* i obtenció dels conjunts de dades. Alguns exemples són l'obtenció de: identificadors de les empreses que pertanyen al *SP500* segons una data donada, notícies, referències de *Wikidata*, cotització d'obertura, cotització de tancament, màxim diari, mínim diari, volum de mercat, dies d'inversió (la borsa tanca en caps de setmana o festius per exemple), etc.

3.2 Modificació dels objectius

Cal esmentar que els objectius del treball s'han vist modificats després d'una anàlisi exhaustiva de l'estat de l'art. Inicialment, es va establir l'objectiu principal de millorar el rendiment (precisió per exemple) dels models més punters actuals. Deixant de banda que és un objectiu altament ambiciós, els models que estan disponibles de codi obert, estan parcialment publicats perquè no puguin ser reconstruïts per complet amb els mateixos resultats de rendiment. Sol mancar la generació de la construcció del graf o els conjunts de dades utilitzats, de manera que impossibilita l'objectiu passat amb el temps disponible per a fer aquest treball de fi de grau. Això és molt comprensible, pel fet que alguns d'aquests models aconsegueixen un retorn anual de més del 60% de la inversió anual, sent eines molt preuades per a generar diners.

Per aquestes raons, s'ha reconstruït una versió capada (que és capaç de batre el mercat) del codi parcialment publicat d'un article punter, per a testejar el rendiment dels grafs utilitzats. Serà més que suficient per als objectius que s'han redefinit per causa que, l'enfocament està situat en la construcció de grafs. Per tant, el rendiment dels grafs podrà ser comparat de manera relativa.

3.3 Requeriments funcionals i no funcionals

A continuació es llisten els requisits funcionals i no funcionals mitjançant el model FURPS (*Functionality, Usability, Reliability, Performance, Supportability*). La primera lletra del següent llistat indica l'atribut i el número identifica l'element dins de cada categoria.

F1: Generar un conjunt de dades d'indicadors tècnics, notícies i dades de la indústria de les empreses del *SP500* com a graf a partir de APIs i *crawlers*.

F2: Que sigui capaç de ser entrenat a partir de diverses metodologies de construcció de grafs de manera flexible i eficaç.

F3: Que el model pugui ser validat per adaptar els hiperparàmetres amb dades diferents de la fase d'entrenament.

F4: Que el model sigui capaç de realitzar inferències amb nous conjunts de dades borsàries (diferents de les que s'ha entrenat i validat el model) sobre la cotització de les accions borsàries emprant el model entrenat (fase de testatge).

F5: Que el model pugui proporcionar mètodes d'avaluació del graf en relació amb un "graf ideal" sense que sigui necessari entrenar-lo.

U1: Que el programari pugui ser controlat mitjançant una interfície de comandes intuïtiva.

U2: Oferir documentació clara i completa de com usar el programari i les seves funcionalitats.

R1: Implementar mecanismes de control i maneig d'excepcions per informar a l'usuari i evitar errors inesperats.

P1: Que el model sigui capaç de batre el mercat i obtingui beneficis anuals.

P2: Que el programari pugui ser executat en un servidor amb GPU *GPX 1080Ti*, 12 *cores* i 2,2 GHz.

S1: Proporcionar suport per a l'escalabilitat del sistema, permetent el processament eficient de grans conjunts de dades.

3.4 Obstacles

- **Millorar l'estat de l'art**: trobar un "graf ideal" i poder realitzar un *benchmark* de rendiment de diferents tipus de construcció de grafs és l'objectiu principal del treball però, també el major obstacle atès que, manquen predecessors d'aquesta línia d'investigació.

- **Batre el mercat**: tot i que s'ha comprovat que les GNN són aplicables al problema de predicció d'accions, suposa un repte no trivial reconstruir un model que sigui capaç de batre el mercat.

- **Inexperiència**: la familiarització amb les GNN és un obstacle, donat que l'autor només té poca experiència amb diferents tècniques d'aprenentatge automàtic però no en GNNs.

- **Grup d'investigació i finances**: l'àmbit de les finances suposa un repte per al grup d'investigació *Barcelona Neural Networking Center (BNN)*, atès que no està especialitzat amb aquest aspecte ni hi ha una línia d'investigació específica dedicada a aquesta aplicació.

3.5 Riscos

- **Gestió del temps**: l'autor del TFG està cursant un grau en economia a la vegada que realitza el TFG. Per aquesta raó, poden sorgir pics d'activitat inesperats que forcin canvis en la planificació del treball.

- **Complexitat de les tècniques algorítmiques**: donat que l'autor té certa inexperiència en l'àmbit de GNNs i són tècniques algorítmiques amb certa dificultat, és possible que requereixi més temps en certes tasques d'implementació del que estava previst inicialment.

- **Accidents**: encara que pugui semblar irrellevant per la baixa probabilitat que succeeixi, s'ha considerat que el risc de patir un accident seria un contratemps important que afectaria la planificació del projecte atès que el temps és molt limitat.

4 Disseny i implementació

4.1 Marc teòric

Aquest punt s'expliquen de manera teòrica i breu el funcionament de les xarxes neuronals que formen part del model.

4.1.1 Unitats recurrents controlades (GRU)

Les unitats recurrents controlades, més conegudes com a *Gated recurrent units*, són un mecanisme de *gating* en xarxes neuronals recurrents. Aquestes van ser introduïdes el 2014 per Kyunghyun Cho [25].

Van sorgir amb l'objectiu de solucionar el problema del gradient minvant, que tenen les xarxes neuronals recurrents (RNN) (classe de xarxes neuronals on les connexions entre estats tenen cicles) [30]. Aquest problema es troba quan s'entrenen xarxes naturals artificials. Aleshores, en cada iteració d'entrenament, els pesos de les xarxes neuronals són actualitzats proporcionalment a la derivada parcial de la funció d'error respecte al pes actual [28]. Per tant, quan la longitud de la seqüència incrementa, el gradient disminueix (o creix incontrolablement), perjudicant la fase d'entrenament [29].

Per solucionar el problema del gradient minvant, la GRU utilitza una porta d'actualització i una porta de reinici. Bàsicament, aquestes són dos vectors que filtren la informació que arriba a la sortida. De manera que poden ser entrenades per conservar informació del passat rellevant per a la predicció. Per tant, la porta d'actualització determina quanta informació del passat ha de passar al futur i la porta de reinici determina quanta informació del passat cal oblidar [30].

Aquests mecanismes són molt beneficiosos en el problema de predicció de cotitzacions d'accions i en general en GNNs temporals, donat que permet captar les dependències al llarg del temps.

La GRU és semblant a una memòria a llarg termini (LSTM), amb el mecanisme de *gating* que permet rebre o oblidar certes *features*. Això no obstant, la GRU no té una portada de sortida, resultant en un mecanisme amb menys paràmetres [27].

El rendiment de les GRUs és semblant al de les LSTMs. Tot i que el mecanisme de *gating* s'ha demostrat avantatgós en general, no s'ha arribat a cap conclusió de quina de les dues unitats de *gating* és millor [26].

A la Figura 7 es pot observar l'arquitectura d'una GRU.

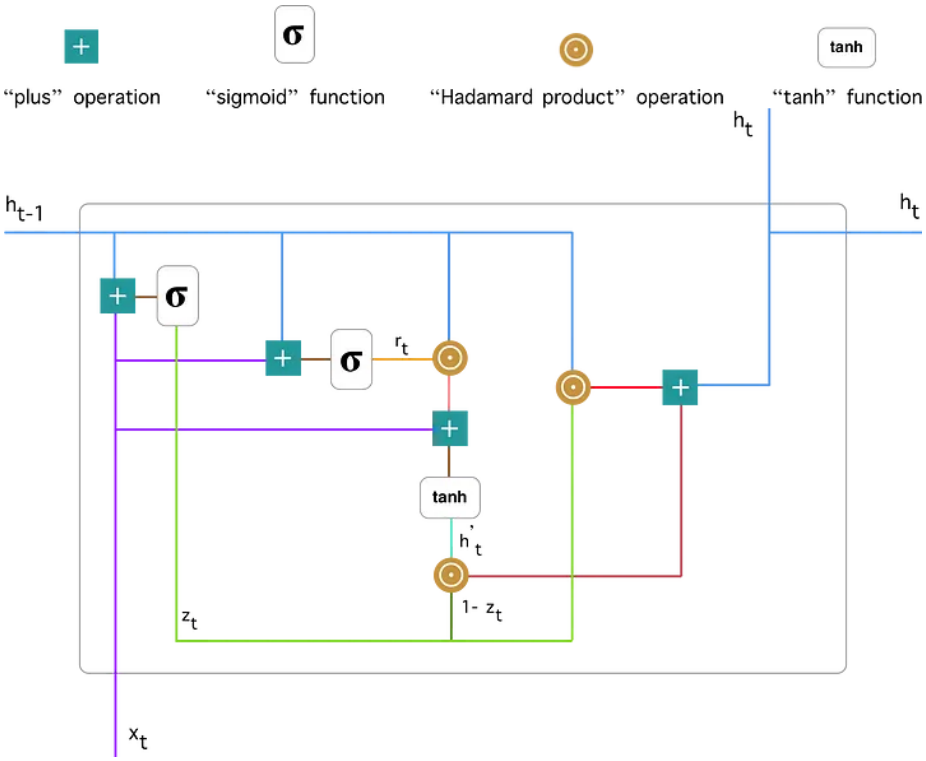


Figura 7: Arquitectura d'una GRU. Font: *Towards Data Science* [30].

4.1.2 *Graph Attention Networks* (GATs)

Les *Graph Attention Networks* (GATs) són un subconjunt de GNN que aprofita els mecanismes d'atenció per a l'aprenentatge de *features* en grafs. Són relativament noves, atès que van ser introduïdes per Veličković el 2018 [31]. Aquestes proposen un mètode que millora com s'agrega la informació dels nodes veïns, en comparació amb les GNNs tradicionals.

En les GNNs estàndard, com per exemple les *Convolutional Networks* (GCNs), els *features* són actualitzats normalment amb la mitjana dels *features* dels nodes veïns. De manera que, no diferencia entre les diferents contribucions dels nodes propers.

Per altra banda, les GATs (vegeu la Figura 8) assignen un coeficient d'atenció a cada node, que indica la importància dels *features* veïns en relació amb l'actualització dels *features* del mateix node. Aquests coeficients amb capacitat

d'aprenentatge són calculats mitjançant un mecanisme d'autoatenció compartit, que calcula l'atenció per cada parell de nodes. Més tard, aquests coeficients són normalitzats en cada "veïnat" utilitzant la funció *SoftMax* [32].

Aquesta alternativa, que permet assignar pesos diferents dels nodes veïns, és més flexible i permet estimular millor el model. Concretament, per al problema de predicció d'accions són molt utilitzades, ja que l'impacte de les empreses veïnes no és homogeni i hi ha empreses que tenen més afectació en altres.

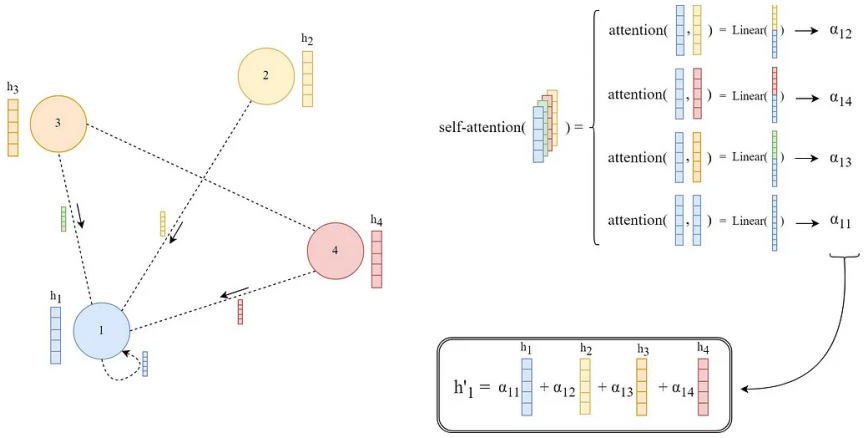


Figura 8: Transmissió del missatge en una GAT. Font: *Towards AI* [33].

4.2 Reconstrucció d'un model existent

Per al disseny i implementació de la GNN s'ha reconstruït un model parcialment publicat en codi obert. L'estructura del model pertany a un article punter anomenat *Temporal and Heterogeneous Graph Neural Network for Financial Time Series Prediction* [16] publicat el 2022, i amb les últimes modificacions de codi publicades el 2023. Abans d'explicar els detalls del model, cal dir que existeixen múltiples diferències entre el model publicat en codi obert i el model descrit a l'article.

El repositori oficial de l'article té un *data leak* [18] en la construcció del graf i, per tant, els mateixos desenvolupadors indiquen que el model més actualitzat de codi obert es troba publicat en un altre repositori [18].

A més, en el model publicat manquen tot el preprocessament de les dades i la generació de les mateixes dades. També manca la construcció dels grafs. Un cop generades aquestes fases, també s'ha hagut de modificar bona part del model per

adequar-se al format de les dades generades, donat que tampoc se sap el format de les dades inexistents per les quals ha estat dissenyat el model inicial. A més s'ha implementat una optimització dels hiperparàmetres, la capacitat d'entrenar per separat el model amb diversos grafs i s'ha redefinit l'arquitectura per poder entrenar també amb un sol graf. Així i tot, després d'investigar múltiples models de codi obert i intentar múltiples reconstruccions, és el model que ha sigut més viable de desenvolupar (ja s'ha explicat al punt 3.1 els interessos existents de no publicar per complet els models).

4.2.1 Diferències entre el codi publicat i l'article

Algunes diferències entre el model publicat (codi) i l'article són:

- ***Encoder***: GRU al codi en lloc d'un Transformer a l'article.
- **Funció objectiu**: regressió lineal per a la variació percentual de les cotitzacions d'accions al codi, en lloc de classificació binària de pujada o baixada a l'article, mitjançant entropia creuada binària.
- **Construcció del graf**: dades de la indústria al codi en lloc de correlacions històriques d'indicadors del mercat a l'article. Encara que, tampoc són publicats els fitxers amb dades de la indústria. Simplement, s'intueix pels noms que assignen a aquests grafs al codi.
- ***Features***: indicadors *alpha* al codi en lloc d'únicament els preus d'obertura de la borsa. Aquests indicadors tampoc estan generats, però estan numerats de manera que poden ser reconstruïts.

4.2.2 Arquitectura del model

Cal esmentar, que la part de generació del graf és la part de l'arquitectura en la qual el treball se centrarà en avaluar. Per tant, la generació del graf s'anirà canviant segons sigui convenient. L'arquitectura del model finalment utilitzat per a la predicció de cotitzacions borsàries té la següent arquitectura (resumida a la Figura 9).

- **Generació del graf**: La primera part és la generació d'un graf. Encara que s'utilitzaran múltiples maneres de generar-lo, l'alternativa proposada inicialment a l'article és a partir de correlacions històriques d'indicadors d'accions. Aquest i altres mètodes de construcció estan explicats amb detall al punt 4.4.

- ***Features* i codificació**: Pel que fa als *features*, s'utilitzen 47 indicadors tècnics diaris diferents publicats a *101 Formulaic Alphas* [19](un recull de 101 indicadors que estan estandarditzats a la indústria de les finances quantitatives) per a cada una de les accions llistades a l'índex SP500. Per a entrenar el model, s'utilitzen finestres dels últims vint dies per a tractar de predir "l'endemà". És a dir, les dades utilitzades per a la predicció d'un dia concret consta d'una matriu tridimensional amb les dimensions (500,47,20) (no són exactament 500 empreses, però això serà explicat més endavant). Aquesta matriu és codificada amb una *Gated recurrent unit* (GRU) per poder modelar les dependències temporals entre accions.

- ***Graph Attention layer***: La tercera part són dues *Graph Attention layers* que calculen la importància dels nodes veïns en el graf de relacions negatives i per separat, en el graf de relacions positives. Per més detall, cal veure el punt 4.1.

- ***Graph Attention layer* heterogènia**: A la quarta part, una capa heterogènia de *Graph attention* que, després de rebre la combinació de representacions, adaptativament calcula la importància i agrega informació de diferents tipus de veïns. Finalment, els valors són normalitzats per a poder fer la predicció del model.

- **Funció objectiu**: El *target* del model són les variacions percentuals diàries de les accions. Un cop fetes les prediccions, s'utilitza un optimitzador Adam [20] per a poder ajustar els paràmetres del model en relació amb la funció de pèrdues. Aquesta està definida com el MSE entre el canvi percentual predit i el canvi percentual real.

- **Estratègia d'inversió pel testatge**: L'estratègia utilitzada per testejar el model, tracta de simular de manera diària la compra del 10% d'accions que més pugen i obrir posicions en curt per al 10% d'accions que més baixen. Més conegut com a *short-selling* en anglès, és una pràctica de venda d'accions, que utilitza un tercer, per a comprar l'acció en una data posterior amb intenció de retornar-li al prestador. D'aquesta manera, s'obtenen beneficis en cas que la cotització de l'acció baixa.

- **Optimització dels hiperparàmetres**: Per a definir els hiperparàmetres s'ha utilitzat la llibreria *Optuna* [49], que està directament desenvolupada per a l'optimització d'hiperparàmetres de models d'aprenentatge

automàtic. S'han realitzat més de 200 proves per poder ajustar els hiperparàmetres (amb una duració de diversos dies d'execució). Per a fer servir *Optuna*, es proporciona un espai d'hiperparàmetres a provar. A continuació, l'eina determina la combinació d'hiperparàmetres que optimitza el model d'aprenentatge automàtic a partir d'una cerca.

La mètrica de rendiment que s'optimitza és el coeficient d'informació (mètrica estandarditzada en la predicció d'accions) en el conjunt de dades de validació. Per evitar esbiaixar l'optimització dels hiperparàmetres amb la inicialització dels paràmetres (que és estocàstica), s'ha optimitzat la mitjana de 5 proves per a cada prova d'hiperparàmetres. D'aquesta manera es redueix la variància i, es dirigeix la cerca amb un mostreig suficient perquè els resultats siguin vàlids. Per altra banda, l'algoritme d'optimització de la llibreria que s'ha usat és el *Tree-structured Parzen Estimator* [48].

- **Períodes de testatge i entrenament**: per predir l'any sencer qualsevol, el model s'entrena amb les dades dels últims 6 anys. L'estudi s'ha focalitzat en l'any 2023 (dades de testatge). Aleshores, l'any 2022 serveix per al conjunt de dades utilitzat per a la validació i, per altra banda, els anys entre 2017 i 2021 (inclosos) serveixen per al conjunt de dades utilitzat per a entrenar el model.

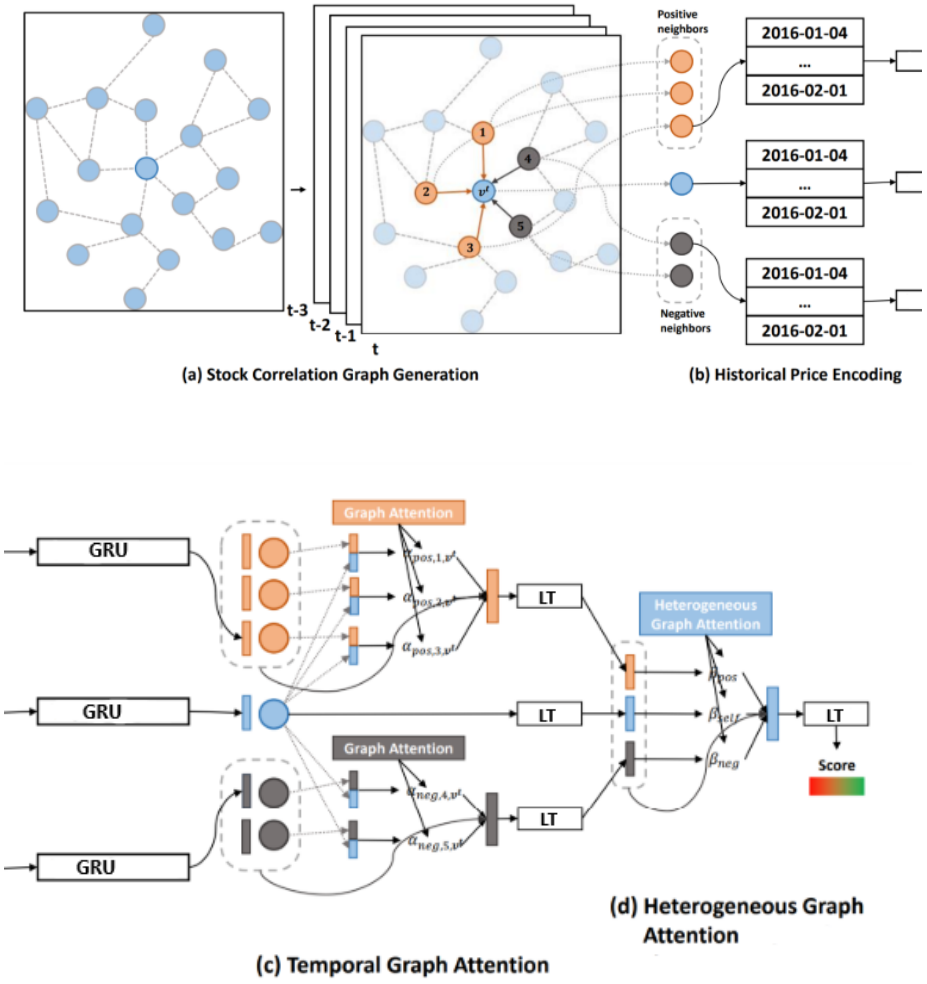


Figura 9: Arquitectura de la GNN heterogènea utilitzada. Font: elaboració pròpia a partir de la figura publicada a *Temporal and Heterogeneous Graph Neural Network for Financial Time Series Prediction* [16]. *LT* a la figura indica *linear transformation*.

4.3 Generació del conjunt de dades i preprocessament

Encara que els fitxers que utilitzen no siguin publicats, els noms que utilitzen en el codi i la lògica intel·ligible del codi permet que puguin ser reconstruïts després d'un estudi exhaustiu del funcionament del codi. El primer pas és l'obtenció de les cotitzacions borsàries de l'índex *SP500*. *Yahoo Finance* és la llibreria

estandarditzada per a obtenir dades financeres. Malgrat això, només es poden obtenir les dades d'identificadors (anomenats *tickers*) d'accions donades. Per aquest motiu, s'ha implementat un *crawler* que obté els identificadors de les empreses del *SP500* mitjançant un llistat de la *Wikipedia* [21]. No obstant això, el *SP500* és un índex dinàmic, atès que varia depenen de quines empreses tenen més capitalització de mercat.

Per tant, si es vol entrenar o testejar amb dades passades és necessari saber quines empreses estaven llistades a l'índex en el seu moment. En cas de no fer-ho estaríem entrenant el model amb informació del futur, això suposa un gran avantatge del qual no es pot disposar per entrenar, ja que aquelles empreses que en el passat tenien perspectives d'entrar a l'índex, són aquelles que més van créixer.

Per aconseguir-ho, s'ha implementat un *crawler* a partir d'una taula de la *Wikipedia* [21] que llista els canvis històrics que hi ha hagut a l'índex. Sabent en quina data una determinada acció va entrar o sortir de l'índex, permet iterar a través de les dates d'inversió per a crear un llistat de *booleans* que ens indiquin si en una data determinada l'acció estava llistada o no.

Les dates d'inversió s'obtenen mitjançant *Yahoo Finance* (la borsa tanca en festius o caps de setmana). Aleshores mitjançant els *crawlers* mencionats anteriorment, s'obtenen les cotitzacions d'obertura, cotitzacions de tancament, volum, màxim i mínim per a cada dia de les dates d'inversió de totes les empreses del *SP500*. Aquestes taules són generades en *DataFrames* de la llibreria de *Python*, *Pandas*. Llavors, s'emmagatzemen en fitxers HDF5, que són utilitzats pel model.

Un cop els indicadors diaris (màxim, mínim, volum, cotització d'obertura i cotització de tancament), són emmagatzemats per a cada una de les dates, per a cada una de les empreses del *SP500*, es generen 47 indicadors tècnics derivats d'aquests indicadors primigenis. S'han utilitzat 47 indicadors dels 101 llistats a *Alpha 101* [19] que donen millors resultats per al model. A continuació hi ha 2 exemples d'aquestes 47 *alphas* per entendre el format:

- **Alpha#6**:

$$-1 \times \text{correlation}(\text{open}, \text{volume}, 10)$$

- **Alpha#13**:

$$-1 \times \text{rank}\left(\text{covariance}\left(\text{rank}(\text{close}), \text{rank}(\text{volume}), 5\right)\right)$$

Pel que fa a les limitacions d'aquestes dades, cal destacar que no existeix cap API gratuïta per obtenir dades d'accions *delisted*. Una acció es considera *delisted* quan és eliminada d'un mercat de valors com el *Nasdaq* o el *New York Stock Exchange* [38]. Alguns dels possibles factors poden ser la decisió de l'empresa de privatitzar-se o la fallida de l'empresa.

4.4 Mètodes de construcció del graf

Els mètodes de construcció del graf escollits són un recull de les metodologies utilitzades a la literatura existent [1], amb una estratègia redefinida pròpia d'aquest treball de fi de grau. S'ha decidit implementar només grafs no dirigits, encara que molts cops les relacions entre empreses sigui dirigida. Aquesta és una pràctica habitual en el desenvolupament de GNNs, per no perdre part de la informació.

4.4.1 Correlació històrica d'indicadors

Aquest mètode de construcció parteix dels següents indicadors diaris obtinguts mitjançant la llibreria de *Yahoo Finance*: màxim, mínim, volum, cotització d'obertura i cotització de tancament. Es genera un graf a partir d'una finestra en la qual es calcula la correlació de cada parell d'empreses. A partir de les correlacions, es generen dues matrius d'adjacències, una matriu amb arestes positives si el llindar de correlació entre accions és major a 0.6 i una altra matriu si el llindar de correlació és menor a -0.6 (valors proposats a l'article).

Per a l'estudi d'aquest mètode de construcció s'han triat 3 modalitats. S'ha testejat amb 3 granularitats temporals per avaluar l'impacte dels diferents tipus de finestres.

- **Correlacions diàries**: per a predir les cotitzacions d'un dia determinat es calculen les correlacions mitjançant la finestra dels últims 20 dies. Aquest mètode és bastant costós computacionalment, atès que s'ha de generar un graf diari per a 7 anys de dades (aproximadament 2555 grafs).
- **Correlacions mensuals**: en aquest cas es genera un graf mensual. Primer, es cerca l'últim dia d'inversió de cada mes, amb la finalitat de calcular la correlació dels últims 20 dies, partint de l'últim dia d'inversió mensual. Es redueixen la quantitat de grafs a generar comparant amb les correlacions diàries, però si s'ha de predir dades de final de mes, el graf està

construït amb dades del mes anterior, la qual cosa comporta una pèrdua d'informació.

- **Correlacions anuals**: en últim lloc, per a predir un any sencer determinat es genera un sol graf de correlacions anuals amb una finestra de 255 dies d'inversió de l'any anterior.

Per a implementar aquests mètodes de càlcul de correlacions s'ha dissenyat per a poder ser calculat amb més d'un procés, donada l'alta demanda de recursos computacionals que comporta.

4.4.2 Propietaris d'accions institucionals

En aquesta construcció del graf, una adjacència existeix quan una empresa és un dels majors inversors institucionals de l'altra. Aquesta informació pot ser complicada de trobar de manera sistemàtica amb dades gratuïtes. No obstant això, es poden trobar els 10 majors inversors institucionals amb *Yahoo Finance*.

Encara que, els majors inversors institucionals poden ser de l'índex *SP500* o no. Això comporta una certa pèrdua d'informació, ja que molts cops aquestes empreses no formaran part del graf. Per altra banda, les dades estan actualitzades a la data 6 de juny de 2023, atès que no hi ha alternatives gratuïtes per a obtenir aquestes dades per a una data donada. Per tant, quan s'utilitzi aquest mètode per a dades d'anys passats, pot suposar una limitació del model que pot ser millorada comprant nous conjunts de dades.

L'obstacle que ha aparegut al desenvolupar aquest mètode és el format amb el qual s'obtenen aquestes dades. *Yahoo Finance* retorna el nom de les empreses, però no retorna l'identificador (*ticker*). Cal tenir en compte, que els noms d'aquestes empreses poden variar lleugerament segons el format. Per exemple, *Blackrock* pot ser retornada simplement com a "*BlackRock*" però també, pot ser escrita com a "*BlackRock, Inc*". Per a resoldre aquest problema, s'ha fet servir la llibreria *Fuzzywuzzy* [34]. Aquesta utilitza la distància Levenshtein per a calcular diferències entre *strings*. Aleshores, partint del *string* retornat per *Yahoo Finance*, s'utilitza la taula del *SP500* de la *Wikipedia* per a poder obtenir el nom de l'empresa llistat que està relacionat amb un identificador específic (després de comparar ambdues *strings* amb *Fuzzywuzzy*).

4.4.3 *Wikidata*

Aquest mètode de construcció utilitza dades de *Wikidata*. *Wikidata* és una base de dades que forma part de la fundació *Wikimedia* (empresa desenvolupadora de *Wikipedia*). Aquesta base de dades és orientada a documents que se centra en *items* que representen un tema, concepte o objecte (a principis de 2023 tenia 1,54 bilions d'*items*) [36]. Cada *item* té un identificador únic i diversos *items* poden tenir certes relacions que també tenen un identificador propi.

A partir d'aquesta font de dades, s'utilitzen dos tipus de relacions seguint l'estratègia utilitzada a l'article *Temporal Relational Ranking for Stock Prediction* [35]. Aquestes poden ser relacions de primer ordre o de segon ordre. Un cop alguna d'aquestes dues relacions sigui identificada, s'assignarà una adjacència entre ambdues empreses (vegeu la Figura 10).

Considerant dues empreses anomenades 'A' i 'B':

- **Relació de primer ordre**: una relació de primer ordre entre 'A' i 'B' significa una relació directa com per exemple *Citigroup* és propietària de *BlackRock*.
- **Relació de segon ordre**: una relació de segon ordre entre 'A' i 'B' es dona quan 'A' té una relació determinada amb una tercera empresa anomenada 'C' i, tanmateix, 'B' té un altre tipus de relació amb 'C'. Per exemple *Boeing* produeix l'avió *Boeing 747* i *United Airlines* l'utilitza.

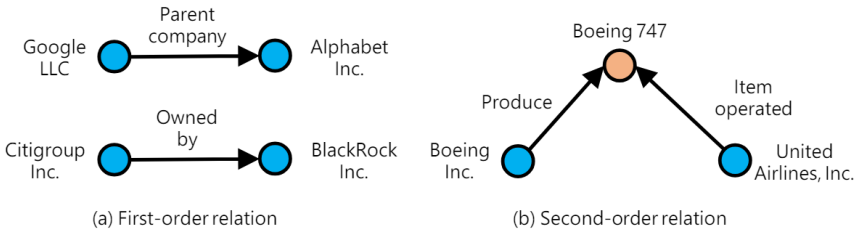


Figura 10: Exemples de relacions de primer i de segon ordre. Font: *Temporal Relational Ranking for Stock Prediction* [35]

Per a desenvolupar aquest mètode s'ha tingut diversos obstacles. Per a obtenir les dades de *Wikidata*, s'ha utilitzat *SPARQL*. *SPARQL* és un llenguatge per a formular *queries* en bases de dades semànticament. Des de desembre

del 2015, *Wikidata* té un punt d'accés que inclou una interfície gràfica molt potent [37].

El primer problema ha sigut relacionar els *tickers* de les empreses amb els identificadors de *Wikidata*. L'inconvenient és que a vegades el subcamp del *ticker* està sota el camp de la borsa, a vegades al revés i a vegades aquests camps no estan enllaçats (vegeu la Figura 11 per a la implementació).

```
query = f"""
    SELECT DISTINCT ?id WHERE {{
        ?id wdt:P31/wdt:P279* wd:Q4830453 .
        {{
            ?id wdt:P249 ?ticker . FILTER(LCASE(STR(?ticker)) =
                LCASE("{ticker}")) .
        }} UNION {{
            ?id p:P414 ?exchangesub .
            ?exchangesub pq:P249 ?ticker . FILTER(LCASE(STR(
            ?ticker)) = LCASE("{ticker}")) .
        }}
        SERVICE wikibase:label {{ bd:serviceParam wikibase:
            language "[AUTO_LANGUAGE],en". }}
    }}
    """
```

Figura 11: *Query* utilitzat per a obtenir els identificadors de *Wikidata*. Font: elaboració pròpia.

Un cop emmagatzemats els identificadors de *Wikidata* en relació amb els *tickers*, s'han obtingut 5 relacions de primer ordre i 51 relacions de segon ordre (vegeu Figura 12) [35]. Les relacions són gairebé les mateixes que les utilitzades a *Temporal Relational Ranking for Stock Prediction* [35], encara que la relació de segon ordre *P361* (part d'un objecte el qual el subjecte és una part) s'ha descartat, atès que aquesta relació implicava el fet d'estar llistat a l'índex *SP500*. Això comportava adjacències per a totes les empreses, per la qual era una relació ineficient. A l'annex F es poden trobar el llistat de totes les relacions utilitzades.

Una limitació d'aquest mètode de construcció és que les dades són actualitzades al present. Això pot portar pèrdues d'informació si s'entrena amb dades de fa molts anys. Encara que, moltes d'aquestes relacions són bastant estables al llarg del temps i, per tant, aquesta ineficiència a priori tampoc hauria de tenir un gran impacte. Algunes d'aquestes relacions tenen a veure amb la indústria, el fundador o cofundador de l'organització, l'empresa propietària, etc.

```
query = f"""
        SELECT DISTINCT ?relationship WHERE {{
            VALUES ?company1 {{ wd:{company_wikidata_id1} }}
            VALUES ?company2 {{ wd:{company_wikidata_id2} }}

            {{
                ?company1 ?relationship ?company2 .
                FILTER(?relationship IN (wdt:P127, wdt:P155,
                    wdt:P156, wdt:P355, wdt:P749))
            }}
            UNION
            {{
                ?company1 wdt:P31 ?instanceProduct .
                ?company2 wdt:P366 ?instanceProduct .
                BIND(wdt:P366 AS ?relationship)
            }}

            # [...] Altres relacions de segon ordre
            UNION
            {{
                ?company1 wdt:P31 ?instanceIndustry .
                ?company2 wdt:P452 ?instanceIndustry .
                BIND(wdt:P452 AS ?relationship)
            }}
        }}
        """
```

Figura 12: *Query* utilitzat per a obtenir les relacions de *Wikidata*. Font: elaboració pròpia.

Un obstacle que es va trobar en implementar aquest mètode de construcció són els límits de *queries*. Es poden obtenir 60 segons de temps de processament cada 60 segons i només es poden tenir 30 *queries* d'error per minut. Tenint en compte que s'ha d'iterar per tots els parells d'empreses, s'obtenia un error de "*too many requests*". Això es va resoldre implementant un mecanisme de control de l'error per esperar el temps necessari per poder fer una altra petició (vegeu la Figura 13).

```
retry_delay = 10
max_retries = 100
retries = 0
    while retries < max_retries:
        try:
            sparql.method = 'POST'
            sparql.setQuery(query)
            sparql.setReturnFormat(JSON)
            # Perform the query
            results = sparql.query().convert()
            relationships =
                [result['relationship']['value'].split('/')[-1]
                for result in results['results']['bindings']]
            return relationships
        except Exception as e:
            print("delay!")
            print(e)
            time.sleep(retry_delay)  # Wait before retrying
            retries += 1
            retry_delay *= 2  # Exponential backoff

    raise Exception("Too many requests!")
```

Figura 13: *Query* utilitzat per a obtenir les relacions de *Wikidata*. Font: elaboració pròpia.

4.4.4 Indústria

Per a obtenir relacions de les empreses mitjançant la indústria a la qual pertanyen s'ha utilitzat el *Global Industry Classification Standard (GICS)*. El *GICS*, és una taxonomia de la indústria desenvolupada el 1999 per *MSCI* i *Standard & Poor's (S&P)* per a l'ús de la comunitat global financera. Consta d'11 sectors, 25 sectors d'indústria, 74 indústries i 163 subindústries (vegeu la Figura 14). En aquesta classificació consten totes les empreses públiques més grans. Això no obstant, és complicat trobar algunes d'aquestes dades de manera sistemàtica i gratuïta, com per exemple, les subindústries.

Al llistat de la *Wikipedia* del *SP500*, es poden trobar les dades del sector GICS i de la subindústria GICS, però només per a les empreses que es troben actualment a l'índex. No hi ha cap altra manera de trobar aquestes dades de manera general i gratuïta mitjançant cap API. Malgrat això, el mètode recomanat en diversos fòrums [40] és la web *FinViz* [39]. Aquesta, proporciona una

classificació semblant a GICS, però cal cercar les empreses cas per cas, atès que els noms de les indústries i subindústries varien.

Aleshores, suposa un increment de feina significatiu, ja que desenvolupar aquesta classificació pot ser costós temporalment. En el cas del treball s'ha optat per desenvolupar manualment la classificació per a les empreses que han entrat i sortit del SP500 des de l'actualitat (maig de 2024) fins a l'inici de 2023. Això suposa una limitació, donat que a l'entrenar el model en anys passats, la informació d'indústria de les empreses que van sortir del SP500 abans del 2023, no ha estat inclosa. Malgrat això, l'any de testatge (2023), sí que compta amb una construcció del graf actualitzada.

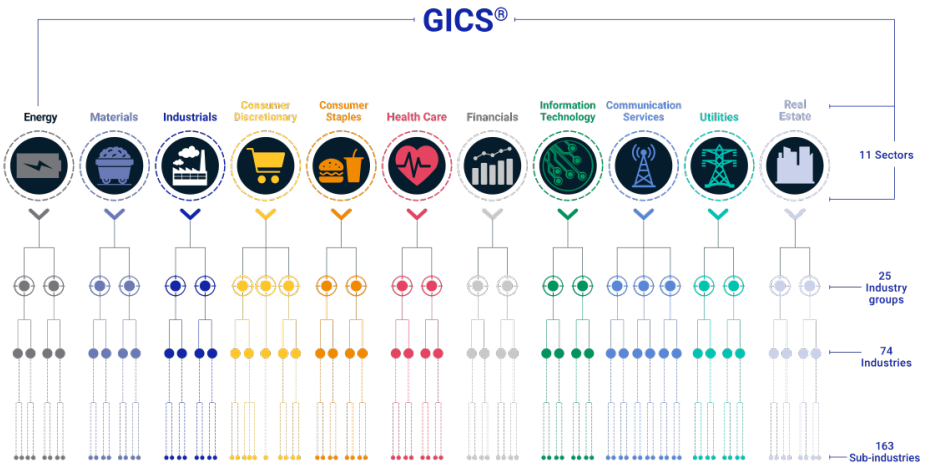


Figura 14: Esquema de la taxonomia d'indústria GICS. Font: fons d'inversió MSCI [41].

Per a construir el graf, s'afegeix una adjacència en cas que dues empreses pertanyen a la mateixa indústria. Concretament, s'han implementat dues versions d'aquest mètode de construcció: una construcció basada en la indústria GICS i una construcció basada en la subindústria GICS. D'aquesta manera també serà possible avaluar l'eficàcia de la granularitat de la indústria.

4.4.5 Notícies

Una informació molt valuosa que pot considerar-se per a la construcció del graf, però també com a *features*, són les notícies de les empreses. Els models més

sofisticats, compten amb mètodes per al processament del llenguatge natural per extreure informació de notícies de manera sistemàtica.

En aquest treball, s'ha volgut incloure un mètode de construcció basat en la informació de notícies, encara que és l'alternativa que més recorregut té per a poder ser millorada en el futur.

Concretament, s'utilitza un conjunt de dades de codi obert molt innovador anomenat *Financial News and Stock Price Integration Dataset (FNSPID)* i publicat el febrer de 2024. Ha sigut dissenyat per a la predicció d'accions amb informació qualitativa i quantitativa. Té preus 29,7 milions d'accions i 15,7 milions de notícies per 4775 empreses del SP500 llistades entre 1999 i 2023 [43] (vegeu la Figura 15).

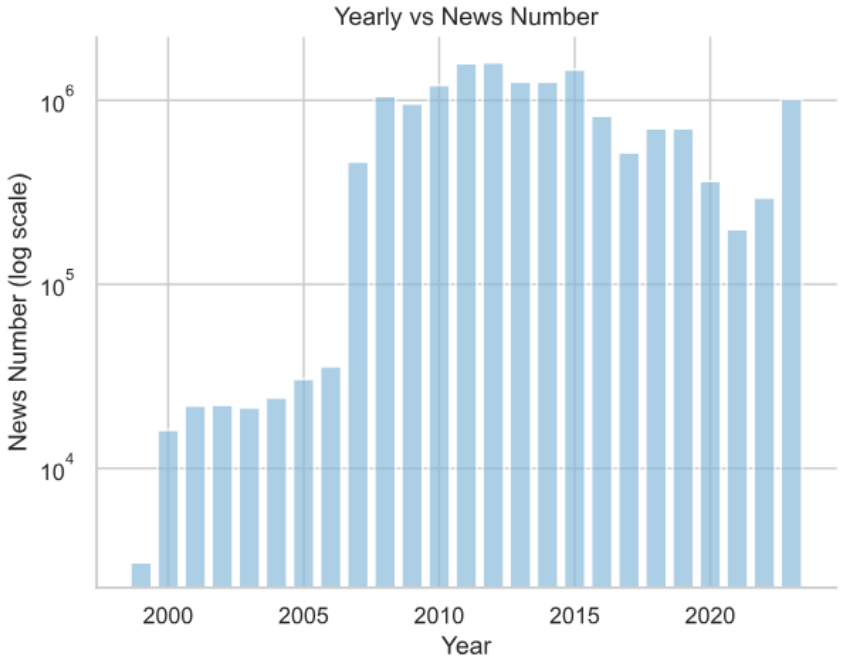


Figura 15: Nombre de notícies al llarg dels anys. Font: *Financial News and Stock Price Integration Dataset* [42].

Aquestes notícies estan extretes de pàgines web de notícies. A l'article on es publica, justifiquen amb diverses proves la qualitat de les dades i expliquen també, com extreure anàlisi del sentiment de compra de les accions basat en

les notícies [42]. Malgrat això, com solament un dels diversos mètodes de construcció del graf utilitzat en aquest treball de fi de grau, s'ha optat per mantenir certa simplicitat. D'aquesta manera, s'afegeix una adjacència en cas que dues empreses han estat esmentades, de forma significativa, en una de les notícies d'un mes determinat. És a dir, s'ha obtingut un graf per a cada mes dels períodes d'entrenament, validació i testatge.

4.5 Graf ideal

Després de moltes proves, s'ha descobert que el model obté uns resultats desproporcionadament bons quan s'entrena el model amb el graf d'adjacències generat amb les correlacions d'indicadors (màxim, mínim, volum, cotització d'obertura i cotització de tancament) del mes sencer actual en comptes d'entrenar amb les correlacions del mes passat. Per la qual cosa, no serveix per a entrenar el model, atès que aquest graf només es pot obtenir amb dades "del futur" (per exemple, el primer dia del mes s'entrena amb un graf creat a partir de correlacions dels 30 dies següents).

Tanmateix, serveix com a "graf ideal" amb el qual, podem mesurar la distància dels diferents mètodes de construcció de grafs. Aleshores, és possible la creació d'un *benchmark* i poder comparar diferents mètodes de construcció, de manera independent al model entrenat.

Cal indicar, que aquest "graf ideal" es pot representar com una matriu binària de dimensions (500,500) (aproximadament). Això podria ajudar a reduir el soroll de les correlacions per a reformular el problema de predicció d'accions com un problema de predicció d'adjacències. Donat que, si el mètode de construcció és capaç d'anticipar quines empreses afectaran altres, el rendiment de la predicció de cotitzacions millorarà.

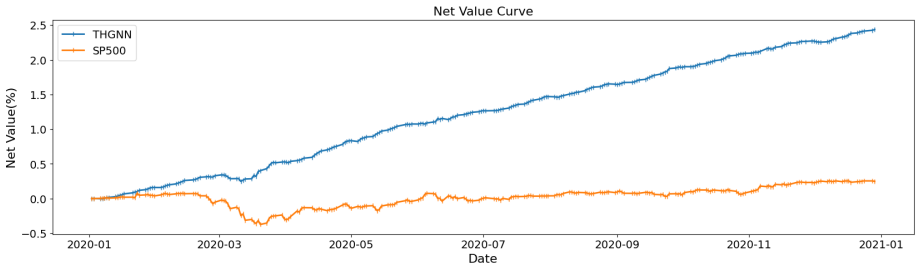


Figura 16: Gràfic del retorn de beneficis segons la data utilitzant el graf "ideal", empreses del *SP500* al 2020. Elaboració pròpia.

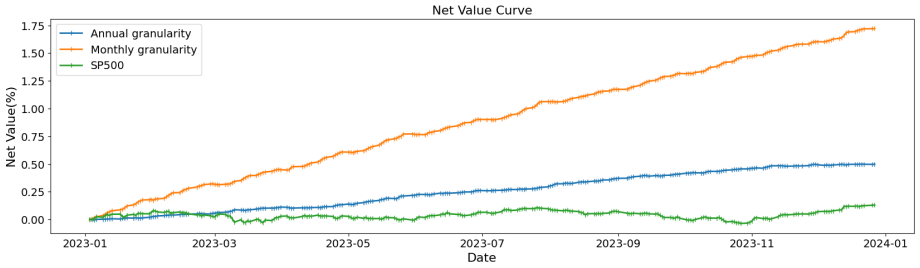


Figura 17: Gràfic del retorn de beneficis segons la data utilitzant el graf "ideal", empreses del *SP500* el 2023. Elaboració Pròpia.

En finances normalment, se sol buscar un retorn de la inversió anual (guanys percentuals esperats) que pugui batre a la inflació. Valors per sobre el 7% són convencionalment considerats bons (tenint en compte el retorn esperat de l'índex *SP500*) [50]. A les Figures 16 i 17 es pot veure clarament com el retorn de la inversió (guanys percentuals esperats) del 175% anual el 2020 i el 250% el 2023 (granularitat mensual), justifiquen l'elecció de la construcció del graf "ideal" (com a referència de comparació). No obstant, s'ha de tenir en compte que la granularitat d'aquests grafs, per poder comparar els altres grafs (estables durant el temps o de diferent granularitat). En el cas de la granularitat anual, el rendiment disminueix del 175% de retorn anual el 2023 al 50%. Això és molt comprensible, atès que en comptes de 12 matrius d'adjacències, el graf anual només en té una que "resumeix" aquestes 12. Per tant, hi ha una pèrdua d'informació, però igualment els resultats són molt bons i justifiquen ambdós

mètodes com a grafs de referència.

A més, cal esmentar la poca volatilitat que tenen aquests grafs de referència. Inclús amb la gran caiguda del 2020 al mes de maig, el graf es manté molt estable aconseguint una tendència de creixement pràcticament constant.

5 Resultats

Encara que l'arquitectura del model estava inicialment dissenyada per a utilitzar dos grafs relacionals, s'ha considerat necessari avaluar també els grafs per separat. Per tant, s'ha modificat l'arquitectura per poder fer una comparació de grafs 1 a 1 i, per altra banda, es compararan tots els parells de grafs per poder dur a terme el *benchmarking* de manera exhaustiva. Cada graf ha sigut testejat 5 cops per mitigar la variància en la inicialització dels paràmetres. Aquestes proves han durat dies en ser executades en un servidor amb GPU *GPX 1080Ti*, 12 *cores* i 2,2 GHz. Les mètriques que s'han utilitzat per avaluar el rendiment dels grafs són:

- **IC**: El coeficient d'informació mesura la correlació entre el preu predit i el preu real. Té un rang de -1 a 1, on 1 indica una correlació perfecta positiva entre el preu de dues accions i -1 indica correlació perfecta negativa. Un major IC implica una capacitat major predictiva del model.
- **Rank IC**: És similar a IC, però utilitza el rànquing dels resultats predits en comparació amb el rànquing dels resultats reals per a calcular la correlació.
- **ARR**: mesura el retorn de guanys anuals (*Annual Rate Return).* Es calcula amb la mitjana dels beneficis diaris multiplicada pels dies d'inversió de tot l'any. No s'ha de confondre amb els beneficis que es mostren als gràfics de retorn anual, donat que aquests beneficis es calculen invertint el 10% d'accions amb més increment esperat diari i el 10% amb més baixada esperada.
- **AV**: la volatilitat mitjana (*Average Volatility*) mesura la variabilitat de les fluctuacions dels retorns de les accions. És una mesura de risc, on major volatilitat implica major incertesa i més variància.
- **Sharpe**: el *Sharpe Ratio* és una mesura de retorn ajustat al risc. Es calcula dividint els retorns per la desviació estàndard de les accions. Un major *Sharpe ratio* indica uns retorns més alts ajustats al risc.
- **WR**: la taxa de guanys (*Win Rate*) també podria ser denominada com precisió (*accuracy*) calcula el percentatge esperat d'inversions que tenen un retorn positiu. Es calcula mesurant quines accions s'encerta que pujaran i també, quines s'encerta que baixaran. Una predicció aleatòria tindria una

Win Rate del 50%. Cal dir que els valors de l'estat de l'art per aquest indicador estan entre el 50% i el 60%. Fluctuacions petites que poden semblar irrellevants en les centèsimes d'aquesta mesura, poden aportar un canvi altament significatiu en els guanys anuals.

- **MDD**: la disminució màxima (*Maximum Drawdown*) mesura quina és la baixada consecutiva més significativa durant l'any. És una mesura de risc on menor *MDD* indica major control de l'estratègia d'inversió quan hi ha caigudes en el mercat o les prediccions no són encertades.

Els resultats dels diferents mètodes de construcció del graf per a l'any 2023 són els següents.

5.1 Resultats dels mètodes de construcció de grafs 1 a 1

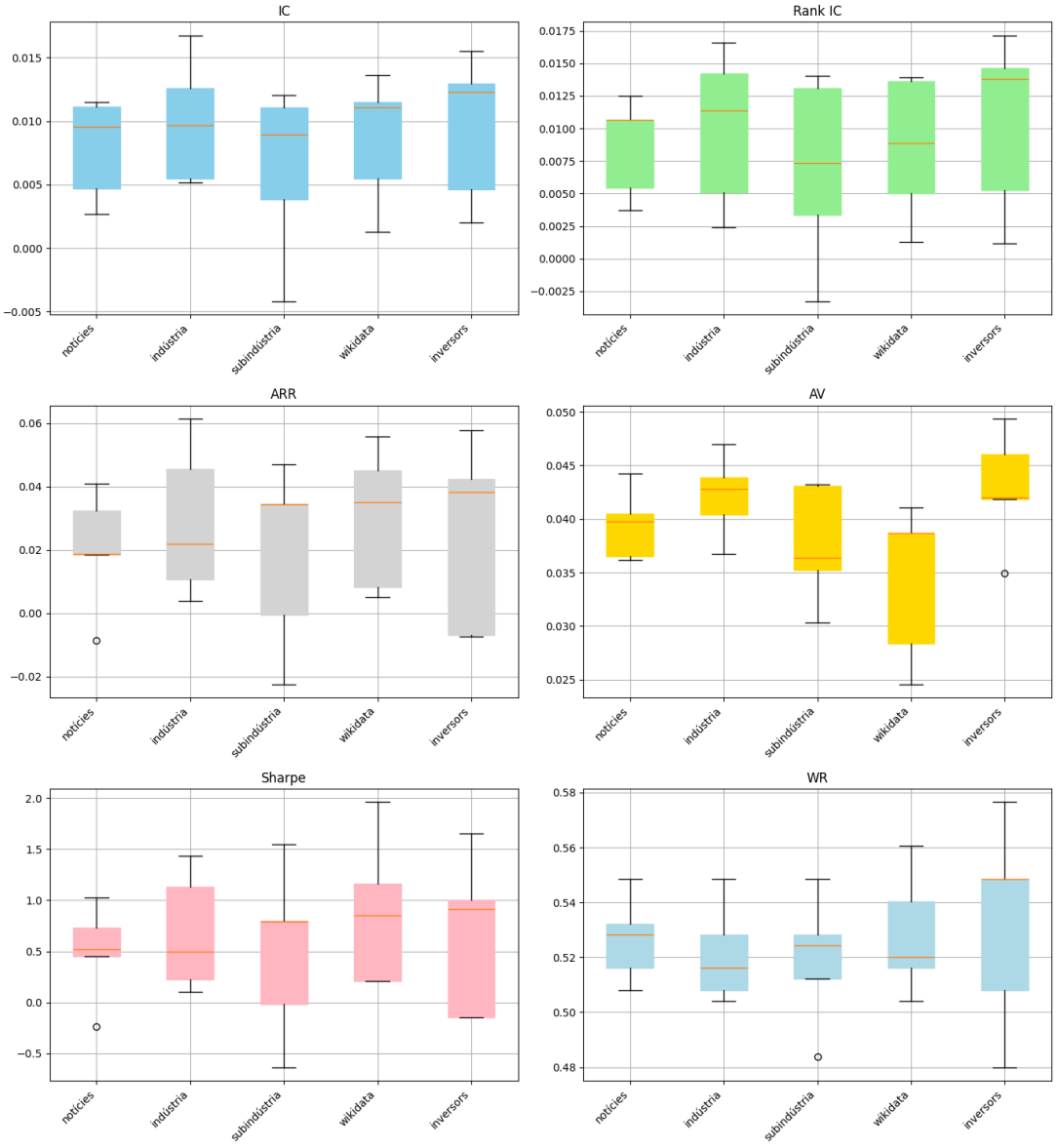


Figura 18: Rendiment de grafs durant l'any 2023 1 a 1. Font: elaboració pròpia.

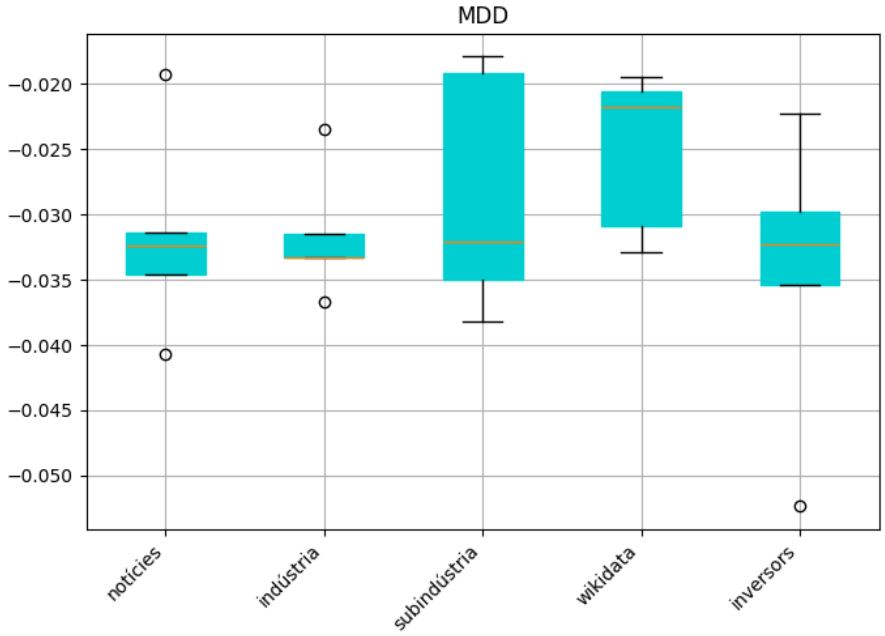


Figura 19: Rendiment de grafs (MDD) durant l'any 2023 1 a 1. Font: elaboració pròpia.

Tal com es pot veure a les Figures 18 i 19, el graf que obté majors beneficis és el graf d'inversors institucionals. Tot i això, podem veure als *boxplots* que aquest graf té una altra variança de resultats que pot ser comprovada amb l'indicador de volatilitat AV i també, a la Figura 19. En aquest cas lidera en retorns però també en risc. Si analitzem el *Sharpe Ratio* (s'ajusta el retorn amb el risc), el graf de *Wikidata* retalla distàncies i es proposa com una de les millors alternatives amb una volatilitat relativament baixa amb el menor valor absolut de MDD.

5.2 Resultats dels mètodes de construcció de grafs 2 a 2

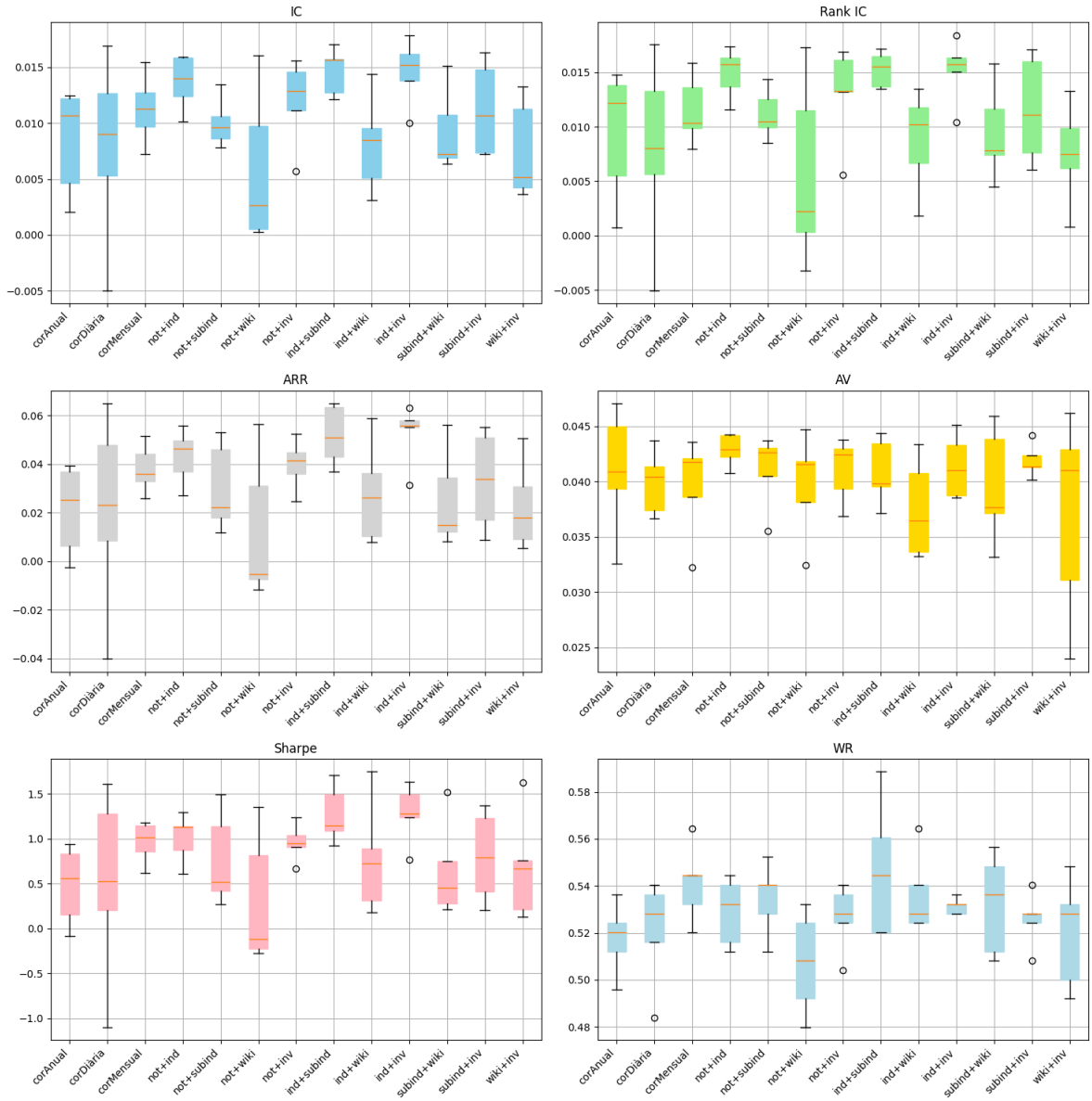


Figura 20: Rendiment de grafs durant l'any 2023 2 a 2. Font: elaboració pròpia.

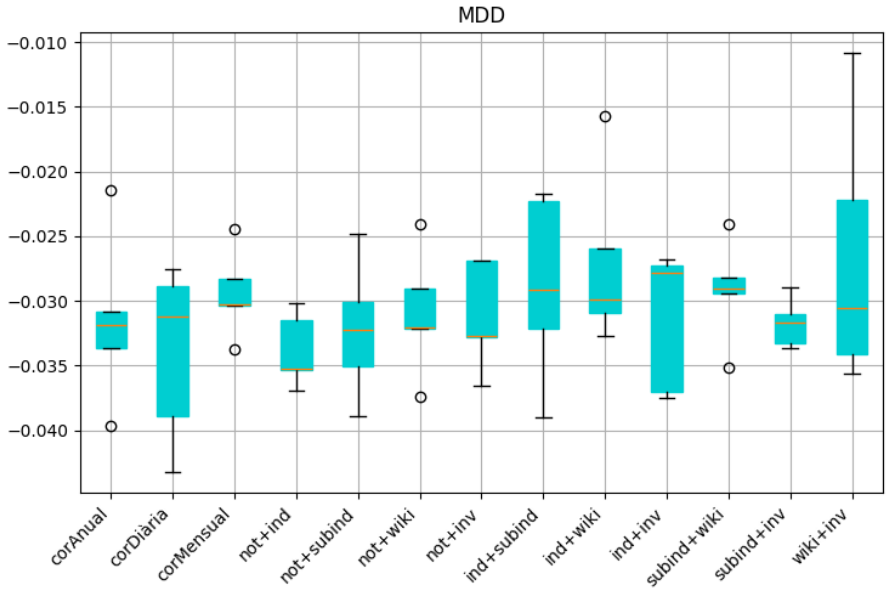


Figura 21: Rendiment de grafs (MDD) durant l'any 2023 2 a 2. Font: elaboració pròpia.

Tot i que el graf de *Wikidata* era dels millors comparant els grafs 1 a 1, en aquest cas 2 a 2 (vegeu les Figures 20 21), n'hi ha de millors. Les més exitoses són les combinacions del graf de construcció d'indústria. Podria semblar que indústria combinada amb el graf de subindústria és informació repetida, però combinant les dues granularitats aconsegueixen la major mitjana d'IC i de WR, seguida de la combinació d'indústria més el graf d'inversors que aconsegueix ser el graf amb més retorns anuals. Si s'observa el *Sharpe ratio*, indústria + inversors continua sent el graf amb millors resultats. També té un dels menors valors absoluts de MDD amb una baixa variància.

Pel que fa a la granularitat, pot ser vist com un augment significatiu de la quantitat de grafs de correlacions (correlacions diàries envers correlacions mensuals) no té per què aportar millors resultats. Tot i passar de 12 grafs de construcció a 255 a l'any, l'augment del soroll repercuteix en un empitjorament dels resultats.

En general, es pot observar una millora significativa en la majoria de les mètriques en l'agregació d'un segon graf. En futurs estudis es podria estudiar quines són les limitacions d'agregar grafs a l'arquitectura per investigar quina

quantitat i qualitat d'informació aporta millors resultats. No està garantit que quanta més informació el rendiment milloraria, atès que pot ser important reduir el soroll.

5.3 Similitud amb el graf ideal

Per començar a analitzar els grafs primer s'ha comptabilitzat el nombre d'arestes (vegeu la Figura 22), ja que és un factor determinant per a calcular la similitud entre grafs.

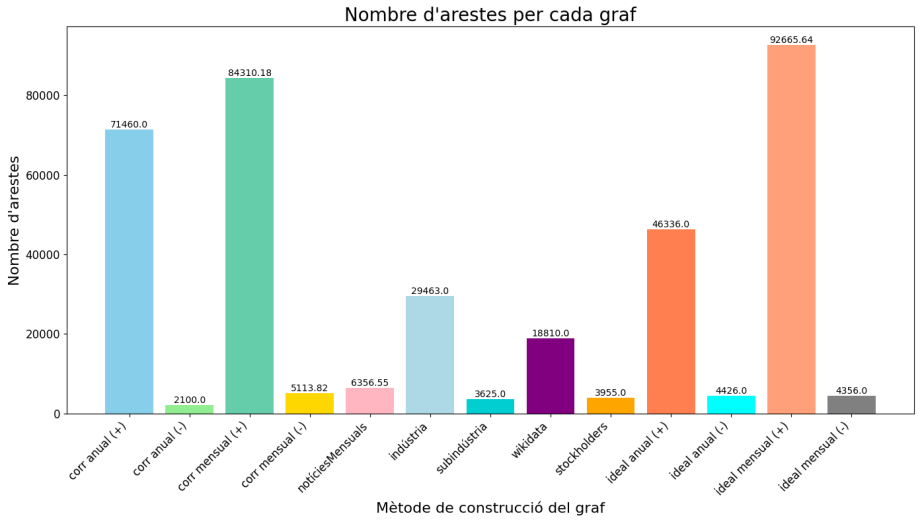


Figura 22: Nombre d'arestes per graf. Font: elaboració pròpia.

Pot ser observat tant com els grafs de correlacions mensuals i anuals de 2022, com els grafs ideals (correlacions mensuals i anuals de 2023) són els que més arestes tenen. Aquesta informació pot ser enganyosa a primera vista perquè cal recordar, que el graf ideal mensual i anual simplement és el graf de correlacions del període següent. Per tant, el graf de correlacions mensuals conté la mitjana de les arestes de gener a novembre i, en canvi, el graf ideal mensual conté la mitjana de febrer a desembre. Els nombres decimals que es poden observar és degut al fet que per a calcular el nombre d'arestes dels grafs mensuals s'ha calculat la mitjana de cada mes.

Per altra banda, cal destacar que l'any 2022 ("corr anual +"), el nombre d'arestes és significativament més alt que el de 2023 ("ideal anual +"). També

pot ser vist que la quantitat d'arestes de correlacions negatives és dels grafs que menys arestes obtenen.

A continuació, s'ha mesurat la similitud entre els grafs i el graf ideal. Per als grafs mensuals s'ha calculat la mitjana de les similituds de cada mes.

Per mesurar la similitud s'han utilitzat les següents mètriques:

- **Similitud Jaccard**: mesura la similitud calculant la mida de la intersecció dels dos grafs, dividida per la mida de la unió. Aquesta mesura s'usa quan la presència d'un element és més important que la seva absència (com és el cas).

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- **Similitud Hamming**: és la inversa de la distància Hamming. Mesura el nombre de posicions diferents entre els conjunts. D'aquesta manera mesura la proporció de posicions entre grafs que són iguals. Aquesta mesura és adient quan tant la presència com absència d'elements és important per igual. Tot i que no és el cas d'aquest treball, s'ha inclòs donat que aporta resultats molt diferents de les altres mètriques de similitud.

$$d(A, B) = \sum_{i=1}^{n} (A_i \oplus B_i)$$

$$H(A, B) = 1 - \frac{d(A, B)}{n}$$

- **Similitud Cosine**: mesura la similitud entre dos grafs codificats com a vectors definits en un espai de producte interior.

$$\text{Cosine}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

- **Similitud Dice**: és similar a la similitud Jaccard però amb una normalització lleugerament diferent. En aquest cas, se li dona més importància a la intersecció.

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

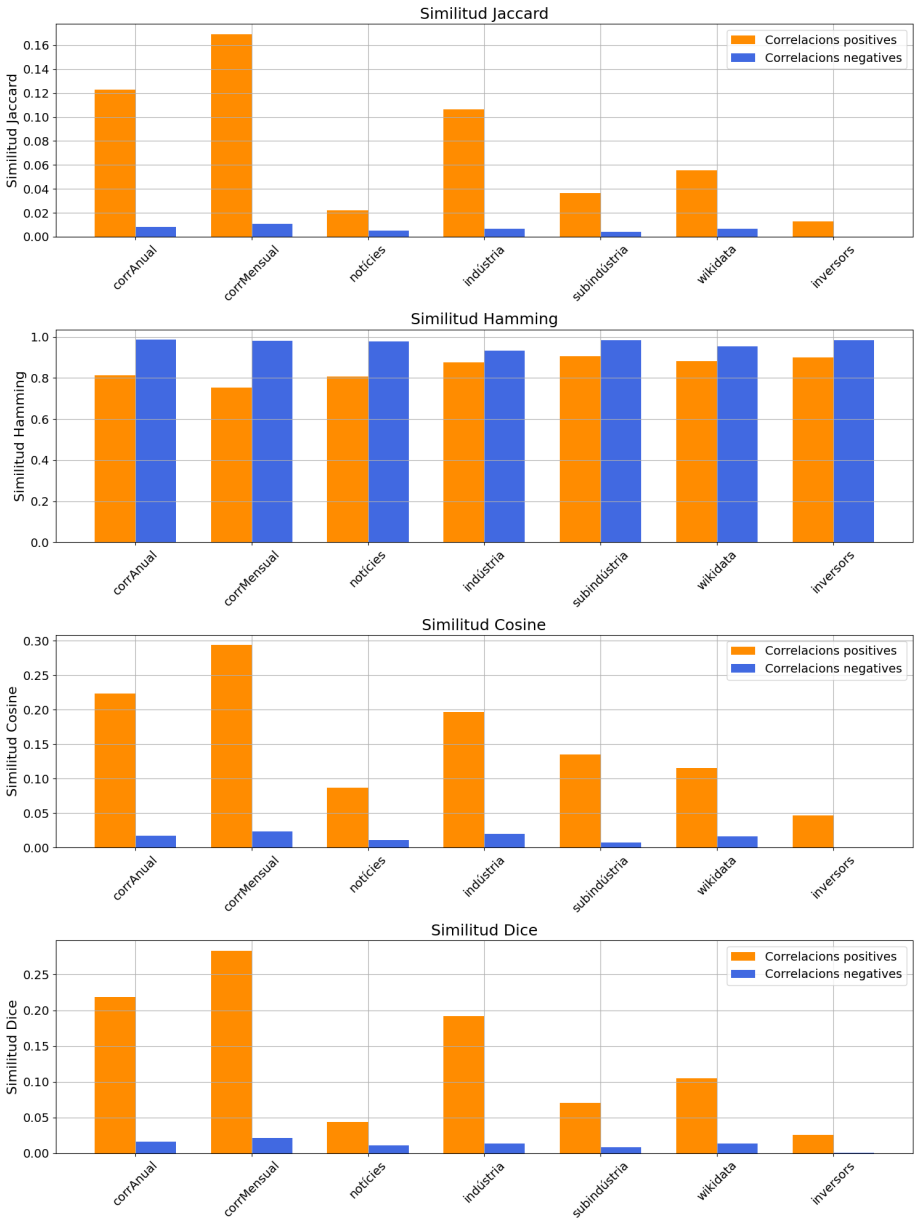


Figura 23: Similitud entre grafs. Font: elaboració pròpia

Analitzant les similituds entre grafs podem extreure les següents observacions. La similitud Hamming té uns resultats molt diferents de les altres atès que dona la mateixa importància als 0s que als 1s. Això premia aquells grafs amb poques adjacències. També es pot veure com la poca quantitat d'adjacències en els grafs ideals de correlacions negatives repercuteix en les similituds, donat que no hi ha cap graf similar (cal obviar els resultats de la similitud Hamming per falta d'adjacències).

Quant als resultats de similitud en relació amb el graf ideal de correlacions positives, es pot veure com la funció de la gràfica és similar en totes les similituds que no són la similitud Hamming. Això corrobora l'evidència de les conclusions i explicacions següents. El graf que és més similar segons els resultats, és el graf de correlacions positives del mes anterior (en comparació amb el del més posterior). Això és comprensible pel fet que és un dels grafs amb més adjacències i moltes de les correlacions d'un mes poden perdurar en el temps per també moure's en la mateixa direcció al mes següent.

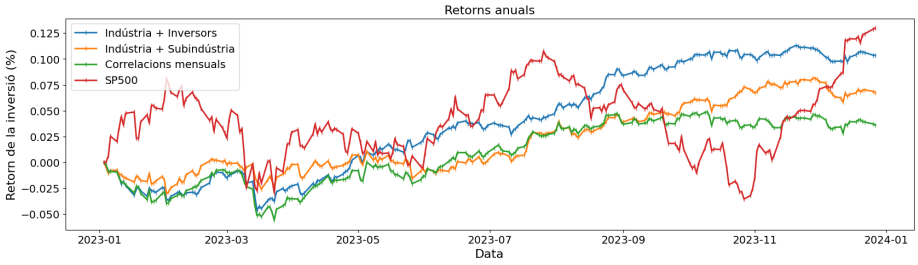
Si més no, a la Figura 22 es pot comptabilitzar una diferència molt més alta d'arestes entre el graf de correlacions mensuals i el graf d'indústria que la diferència proporcional de les similituds trobades. Per tant, podem concloure que no només la quantitat d'adjacències determina la similitud entre aquests dos grafs. Sinó que, la indústria, tenint un terç de les adjacències del graf de correlacions positives mensuals, aconsegueix ser el segon graf amb més similitud amb una diferència proporcional al graf de correlacions molt més baixa que el nombre d'adjacències.

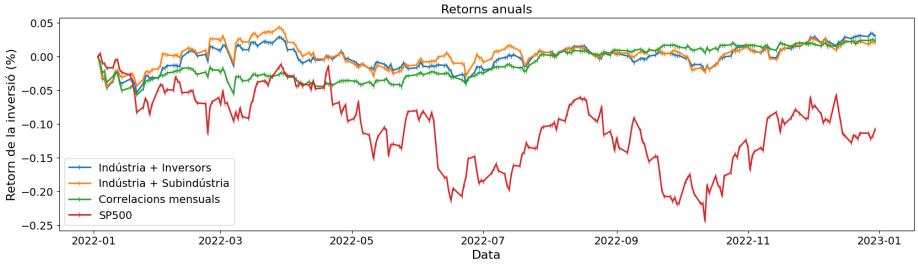
Recordant els resultats trobats a la Figura 20, es pot relacionar el graf amb majors combinacions guanyadores a la resta de grafs, amb un graf amb una major similitud al graf ideal. El graf d'indústria és un dels grafs amb millor rendiment tal com s'ha explicat anteriorment, i és el graf amb millors resultats de similitud (deixant de banda les correlacions del mes anterior). Per altra banda, encara que el graf de correlacions mensuals sigui el millor segons les similituds (atès que es construeixen amb el mateix mètode i hi ha empreses en les quals les relacions perduren durant el temps), també és un dels grafs que té millors resultats de rendiment (Figura 20). En el cas de *Wikidata*, obté el tercer lloc en aquesta anàlisi de similituds corroborant l'alt rendiment en la comparació de grafs 1 a 1 (Figura 18).

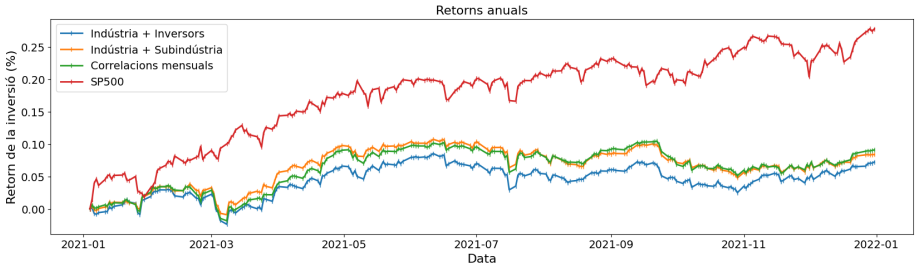
Això no obstant, no hi ha prou evidència per concloure que existeix una correlació directa perfecta entre una major similitud i un major rendiment. Si més no, els resultats sí que poden deixar intuir algun tipus de relació que potser

junt amb altres eines, podrien esdevenir una eina per a la investigació del graf.

5.4 Rendiment durant els anys







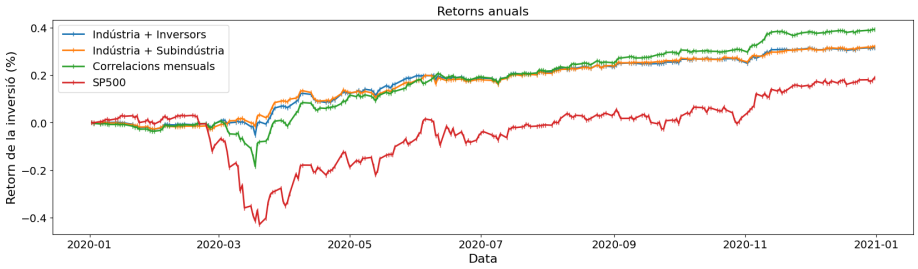


Figura 24: Gràfic del retorn amb diversos grafs de construcció, 2024-2020. Font: elaboració pròpia.

En primer lloc, cal dir que els grafs de construcció que millor han funcionat per a 2023 tenen algunes limitacions (com ja s'ha comentat anteriorment). Les dades dels 10 majors inversors institucionals d'una empresa és extreta durant l'any 2023 i no hi ha manera d'obtenir aquesta informació dels anys anteriors de manera gratuïta. Per altra banda, les dades de subindústria també són per a les empreses que han passat pel SP500 fins a 2023 (vegeu punt 4.4).

En els resultats de la Figura 24 s'ha utilitzat una estratègia d'inversió on cada dia s'inverteix en el 10% d'accions amb més pujada esperada, i s'obren posicions a la baixa per al 10% amb més baixada esperada. Cal diferenciar aquest mètode, de l'indicador d'ARR que s'ha pogut esmentar en punts anteriors. Per altra banda, cal mencionar que el *SP500* que indiquen els grafs, no és el *SP500* que es sol mostrar popularment. Els gràfics del *SP500* més populars estan obtinguts a partir de distribuir el pes en funció de la capitalització de les empreses. En canvi, en aquest cas, el SP500 representa una mitjana de totes les empreses de l'índex sense diferència de pes capitalització de mercat.

Es pot concloure que l'algoritme és guanyador durant el pas dels anys. L'any que obté un major rendiment és durant el 2020, amb retorns anuals de fins al 40%. En canvi, el 2021 aconsegueix un retorn al voltant del 7% (per sota de les fluctuacions del *SP500*). Malgrat això, el 2022, en una tendència a la baixa del mercat aconsegueix mantenir-se estable amb un retorn lleugerament positiu. Per tant, tot i no aconseguir uns retorns anuals enormes el 2021 en una tendència alcista, és capaç d'aguantar la volatilitat i turbulències del 2022 sense pèrdues. El 2023, es pot veure com el graf d'indústria + inversors aconsegueix un retorn considerablement major als altres mètodes de construcció amb un retorn d'entorn el 10%.

També, pot ser dit que l'algoritme té una estabilitat considerable, atès que no es pot percebre una gran volatilitat. Es pot identificar com, al llarg dels anys, l'algoritme té una volatilitat més baixa que la del SP500, assumint menys risc.

6 Conclusions

A tall de conclusió, s'ha aconseguit reconstruir un model que bat al mercat a partir d'una arquitectura puntera basada en una *Graph Neural Network* temporal heterogènia [16]. Aleshores, s'ha analitzat exhaustivament dels diferents tipus de mètodes de construcció de grafs tenint en compte dels factors que afecten les cotitzacions d'accions.

En conseqüència, s'ha desenvolupat un *benchmark* de la qualitat dels grafs relacionals d'empreses per tractar d'avançar en una línia d'investigació en la qual mancaven precedents (vegeu 1.4). S'ha assolit mitjançant estratègies de construcció pròpies, que deriven de diferents mètodes de construcció existents a l'estat de l'art. S'ha desenvolupat mitjançant la implementació de *crawlers* on s'ha obtingut relacions d'indústria, estructura de l'empresa, subministrament, inversors institucionals, notícies, correlacions d'indicadors, etc. També han estat documentades totes les estratègies i desenvolupament del codi, justificats amb gran quantitat de fonts d'articles. A més, tot el codi serà publicat al repositori https://github.com/mimugara.

Les combinacions amb millors resultats han estat les combinacions d'indústria + subindústria i la combinació d'indústria + inversors. Encara que la construcció del graf mitjançant notícies no és de les millors, té molt potencial per a ser millorada amb tècniques més avançades del processament del llenguatge natural.

D'aquesta manera no només s'han avaluat els mètodes de construcció mitjançant múltiples mètriques financeres, sinó que també s'ha proposat una mesura de similitud amb un "graf ideal". Aquesta avaluació originària d'aquest treball de fi de grau, podria permetre optimitzar el desenvolupament per a possibles futures investigacions en la construcció de grafs sense que sigui necessari entrenar i testejar el model. Això pot ser molt eficient en problemes com la predicció de cotització d'accions, atès que el gran volum de dades requereix una capacitat computacional en la qual molts cops s'alenteix el procés de recerca. Encara que, també pot ser aplicat per a multitud de problemes financers en la que sigui necessari modelar les relacions entre empreses com a graf relacional.

En futures investigacions, el "graf ideal" podria ser utilitzat per veure si pot ser aplicat per a un problema de predicció d'adjacències. S'intueix que podria ser beneficiós per a reduir el soroll, i tractar de predir les relacions entre empreses. Un cop obtingudes aquestes relacions, seria possible que milloressin el rendiment de la predicció d'accions borsàries i, inclús altres problemes en el

camp de les finances.

Bibliografia

[1] Wang, J., Zhang, S., Xiao, Y., & Song, R. (2022). *A Review on Graph Neural Network Methods in Financial Applications* (arXiv:2111.15367). arXiv. http://arxiv.org/abs/2111.15367

[2] John McCarthy (1956), The Dartmouth Conference.

[3] Herman, D., Googin, C., Liu, X., Galda, A., Safro, I., Sun, Y., Pistoia, M., & Alexeev, Y. (2022). *A Survey of Quantum Computing for Finance* (arXiv:2201.02773). arXiv. http://arxiv.org/abs/2201.02773

[4] Titcomb, J. (2023). *Supercomputer makes calculations in blink of an eye that take rivals 47 years. The Telegraph.* https://www.telegraph.co.uk/business/2023/07/02/google-quantum-computer-breakthrough-instant-calculations/

[5] Longa, A., Lachi, V., Santin, G., Bianchini, M., Lepri, B., Lio, P., Scarselli, F.,& Passerini, A. (2023). *Graph Neural Networks for temporal graphs: State of the art, open challenges, and opportunities* (arXiv:2302.01018). arXiv. https://doi.org/10.48550/arXiv.2302.01018

[6] Matsunaga, D., Suzumura, T., & Takahashi, T. (2019). *Exploring Graph Neural Networks for Stock Market Predictions with Rolling Window Analysis* (arXiv:1909.10660). arXiv. https://doi.org/10.48550/arXiv.1909.10660

[7] Ying, X., Xu, C., Gao, J., Wang, J., & Li, Z. (2020). *Time-aware Graph Relational Attention Network for Stock Recommendation. Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2281–2284. https://doi.org/10.1145/3340531.3412160

[8] Li, W., Bao, R., Harimoto, K., Chen, D., Xu, J., & Su, Q. (2020). *Modeling the Stock Relation with Graph Network for Overnight Stock Movement Prediction. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 4541–4547. https://doi.org/10.24963/ijcai.2020/626

[9] Sawhney, R., Agarwal, S., Wadhwa, A., & Shah, R. R. (2020). *Deep Attentive Learning for Stock Movement Prediction From Social Media Text and Company Correlations.* In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 8415–8426). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.676

[10] Sawhney, R., Khanna, P., Aggarwal, A., Jain, T., Mathur, P., & Shah, R. R. (2020). *VolTAGE: Volatility Forecasting via Text Audio Fusion with Graph Convolution Networks for Earnings Calls.* In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 8001–8013). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.643

[11] Liou, Y.-T., Chen, C.-C., Tang, T.-H., Huang, H.-H., & Chen, H.-H. (2021). *FinSense: An Assistant System for Financial Journalists and Investors.* Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 882–885. https://doi.org/10.1145/3437963.3441704

[12] Goldman Sachs (2017). *Top of Mind, Global Macro Research* https://www.gsam.com/content/dam/gsam/pdfs/sg/en/commentary/GS_Top%20of%20Mind_June.pdf?sa=n&rd=n

[13] *Salario para Programador en España—Salario Medio. (n.d.). Talent.com.* Retrieved 11 March 2024, from https://es.talent.com/salary

[14] *4 Big Risks of Algorithmic High-Frequency Trading. (n.d.). Investopedia.* Retrieved 11 March 2024, from https://www.investopedia.com/articles/markets/012716/four-big-risks-algorithmic-highfrequency-trading.asp

[15] *Gantter.* (n.d.). Retrieved 10 March 2024, from https://www.gantter.com/

[16] Xiang, S., Cheng, D., Shang, C., Zhang, Y., & Liang, Y. (2022). *Temporal and Heterogeneous Graph Neural Network for Financial Time Series Prediction.* Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 3584–3593. https://doi.org/10.1145/3511808.3557089

[17] Cheng, D., Yang, F., Xiang, S., & Liu, J. (2022). *Financial time series forecasting with multi-modality graph neural network.* Pattern Recognition, 121, 108218. https://doi.org/10.1016/j.patcog.2021.108218

[18] *Finint/THGNN.* (2024). [Python]. Financial Intelligence Lab. https://github.com/finint/THGNN (Original work published 2023)

[19] Kakushadze, Z. (2016). *101 Formulaic Alphas* (arXiv:1601.00991). arXiv. https://doi.org/10.48550/arXiv.1601.00991

[20] Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochastic Optimization* (arXiv:1412.6980). arXiv. https://doi.org/10.48550/arXiv.1412.6980

[21] *List of S&P 500 companies.* (2024). In Wikipedia. https://en.wikipedia.org/w/index.php?title=List_of_S%26P_500_companies&oldid=1224782662

[22] Thorp, E. O. (2017). *A Man for All Markets: From Las Vegas to Wall Street, How I Beat the Dealer and the Market.* Random House Publishing Group.

[23] Quantitative analysis (finance). (2024). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Quantitative_analysis_(finance)&oldid=1223279602

[24] *Intuition vs. Computation: A Brief History of Quant Investing* — BMO Global Asset Management. (n.d.). Institutional CA-EN. Retrieved 21 May 2024, from https://institutional.bmogam.com/ca-en/insights/intuition-vs-computation-a-brief-history-of-quant-investing/

[25] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation* (arXiv:1406.1078). arXiv. https://doi.org/10.48550/arXiv.1406.1078

[26] Gruber, N., & Jockisch, A. (2020). *Are GRU Cells More Specific and LSTM Cells More Sensitive in Motive Classification of Text?* Frontiers in Artificial Intelligence, 3. https://doi.org/10.3389/frai.2020.00040

[27] Gated recurrent unit. (2024). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Gated_recurrent_unit&oldid=1223156982#cite_note-gruber_jockisch-7

[28] Basodi, S., Ji, C., Zhang, H., & Pan, Y. (2020). *Gradient amplification: An efficient way to train deep neural networks.* Big Data Mining and Analytics, 3(3), 196–207. https://doi.org/10.26599/BDMA.2020.9020004

[29] Vanishing gradient problem. (2024). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Vanishing_gradient_problem&oldid=1222680571

[30] Kostadinov, S. (2019, November 10). *Understanding GRU Networks.* Medium. https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be

[31] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph Attention Networks* (arXiv:1710.10903). arXiv. https://doi.org/10.48550/arXiv.1710.10903

[32] Karami, F. (2023, July 20). *Understanding Graph Attention Networks: A Practical Exploration.* Medium. https://medium.com/@farzad.karami/understanding-graph-attention-networks-a-practical-exploration-cf033a8f3d9d

[33] Pichka, E. (2023, July 26). *Graph Attention Networks paper explained with illustration and PyTorch implementation.* Medium. https://pub.towardsai.net/graph-attention-networks-paper-explained-with-illustration-and-pytorch-implementation-eb35edba562c

[34] fuzzywuzzy: Fuzzy string matching in python (0.18.0). (n.d.). [Python]. Retrieved 8 June 2024, from https://github.com/seatgeek/fuzzywuzzy

[35] Feng, F., He, X., Wang, X., Luo, C., Liu, Y., & Chua, T.-S. (2019). *Temporal Relational Ranking for Stock Prediction.* ACM Transactions on Information Systems, 37(2), 1–30. https://doi.org/10.1145/3309547

[36] Wikidata. (2024). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikidata&oldid=1225308942

[37] Wikidata:Servei de consultes SPARQL - Wikidata. (n.d.). Retrieved 8 June 2024, from https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/ca

[38] Salvucci, J. (2023, August 8). What Is Delisting & How Does It Happen to a Stock? TheStreet. https://www.thestreet.com/dictionary/delisting-delisted

[39] FINVIZ.com—Stock Screener. (n.d.). Retrieved 9 June 2024, from https://finviz.com/

[40] Where can I find a full list of all companies classified under GICS (Global Industry Classification Standard) listing their industry and ... (n.d.). Quora. Retrieved 9 June 2024, from https://www.quora.com/Where-can-I-find-a-full-list-of-all-companies-classified-under-GICS-Global-Industry-Classification-Standard-listing-their-industry-and-sub-sector

[41] GICS®—Global Industry Classification Standard. (n.d.). Retrieved 9 June 2024, from https://www.msci.com/our-solutions/indexes/gics

[42] Dong, Z., Fan, X., & Peng, Z. (2024, February 9). FNSPID: A Comprehensive Financial News Dataset in Time Series. arXiv.Org. https://arxiv.org/abs/2402.06698v1

[43] Dong, Z. (2024). Zdong104/FNSPID_Financial_News_Dataset [Jupyter Notebook]. https://github.com/Zdong104/FNSPID_Financial_News_Dataset (Original work published 2024)

[44] Supply and demand. (2024). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Supply_and_demand&oldid=1216987890

[45] Stock market. (2024). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Stock_market&oldid=12

[46] Stock Market. (n.d.). Retrieved 12 June 2024, from https://education.nationalgeographic.org/resource/stock-market

[47] g4dn.2xlarge pricing and specs—Vantage. (n.d.). Retrieved 13 June 2024, from https://instances.vantage.sh/aws/ec2/g4dn.2xlarge

[48] Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. Advances in Neural Information Processing Systems, 24. https://proceedings.neurips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html

[49] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework* (arXiv:1907.10902). arXiv. https://doi.org/10.48550/arXiv.1907.10902

[50] Birken, E. G. (2022, September 28). Return On Investment (ROI). Forbes Advisor. https://www.forbes.com/advisor/investing/roi-return-on-investment/

A Metodologia logística

El treball està plenament enfocat a la investigació d'aprenentatge automàtic en l'aplicació en finances. Atès que és un camp altament complex on és complicat identificar patrons, cal una metodologia molt flexible que permeti implementar algoritmes, provar-los i que puguin ser ràpidament descartats en cas d'identificar mals resultats.

Per aquestes raons, se seguirà la metodologia àgil. Setmanalment, els participants del projecte es reuniran i s'analitzarà el compliment dels objectius. Depenent de les circumstàncies, amb l'ajuda dels indicadors, es faran les modificacions necessàries en la planificació del projecte o l'especificació de les tasques per poder adaptar el treball ràpidament en funció dels resultats obtinguts.

Per altra banda, com que el director i el codirector són experts en el camp de les GNNs podran ajudar ràpidament a l'autor del TFG davant de petites dificultats que puguin sorgir sense haver d'esperar a la següent reunió setmanal. Això facilitarà un accelerament en el ritme de treball i en el compliment de la planificació establerta inicialment.

A part de les reunions setmanals per fer el seguiment s'utilitzarà l'eina *Gantter* [15] per a la planificació de les tasques. El diagrama de Gantt servirà per realitzar el control de les tasques previstes i anirà sent actualitzat dinàmicament davant dels imprevistos que sorgeixin. També s'obrirà un repositori a *GitHub* que permetrà tenir una còpia de seguretat a l'equip i també facilitarà que es pugui fer un seguiment de les actualitzacions del codi de manera eficient i ràpida. No s'ha cregut oportú fer ús d'eines més complexes com Jira perquè no s'ajusta a la casuística de l'equip. El temps del qual es disposa és molt limitat i suposa una corba d'aprenentatge molt pronunciada que podria suposar més obstacles que beneficis en l'agilització del projecte.

B Planificació

Amb l'objectiu d'acabar el treball de fi de grau en el termini establert per la facultat i fer-ho assolint els propòsits del treball, aquesta secció tracta de definir una planificació temporal del projecte dividint-lo en tasques.

Donat que la FIB estima que la càrrega de treball són aproximadament 30 hores per crèdit i el TFG consta de 18 crèdits, el desenvolupament del projecte tindrà una durada aproximada de 540 hores. El començament del projecte (junt amb l'inici del quadrimestre de primavera) va ser el dia 12 de febrer i

el dia límit per acabar-lo ha de ser el 25 juny (considerant la data de lectura més propera tot i que encara no ha estat assignada). Atès que tracta d'un projecte d'investigació dins del grup de recerca *Barcelona Neural Networking Center* (BNN), no hi ha restriccions temporals afegides de cap client o altres circumstàncies extraordinàries.

B.1 Recursos necessaris

Pel que fa als recursos necessaris per a desenvolupar el projecte, calen esmentar els següents requeriments de personal:

- El director del projecte és l'encarregat de dirigir i coordinar l'equip. També, orienta per definir l'abast del projecte i supervisa les tasques que es realitzen setmanalment per conduir el TFG davant de possibles imprevistos.
- El codirector del projecte és l'encarregat d'orientar i ajudar al programador en possibles obstacles en el qual el programador li pugui convenir certa ajuda en tasques específiques i tècniques. Junt amb el director, també supervisa i ajuda a definir l'abast del projecte.
- El programador és l'encarregat de desenvolupar el model i la construcció dels grafs.

Per altra banda, cal esmentar el següent material imprescindible:

- Tres ordinadors per a cada un dels treballadors en el projecte.
- Software: Només es preveu utilitzar *software* gratuït: *Google Scholar i ArXiv* (per accedir a la literatura de l'estat de l'art i els algoritmes necessaris), *Yahoo Finance* i *WikiData* (per als *datasets*), *Visual Studio Code* (edició de codi), *GitHub* (còpia de seguretat i seguiment del codi), *PyTorch Geometric Temporal(PyGT)*, i *Python* (llenguatge de programació).

B.2 Descripció de les tasques

En aquest apartat, es detallen les taques necessàries per a realitzar aquest treball de fi de grau exitosament. Cada una de les tasques és identificada amb un codi, una explicació i una estimació d'hores requerides per a complir-la.

B.2.1 Gestió del projecte

La gestió del projecte esdevé fonamental per a dur a terme totes les tasques administratives. L'assignatura obligatòria GEP de la Facultat d'Informàtica de Barcelona (FIB) forma a l'estudiant per a poder gestionar el projecte de fi de grau i també futurs projectes professionals. Sumant la dedicació de les tasques de gestió del projecte descrites a continuació, obtenim un total de 120 hores.

- **Contextualització i abast (GP1)**: en aquesta tasca es defineixen i s'acoten els objectius del projecte. S'estudia la rellevància del tema en qüestió per a justificar el perquè de l'elecció. També s'indiquen els mitjants amb els quals es desenvoluparà el projecte. És necessari haver fet un estudi de l'estat de l'art per a poder avaluar la rellevància del projecte. La duració de la tasca ha estat de 30 hores.

- **Planificació (GP2)**: en aquesta tasca es defineixen els recursos necessaris i la descripció de les fases i tasques del projecte. S'estudien les dependències de les tasques per a poder establir una planificació que sigui coherent amb les dates límit i requisits del projecte. També s'esmenten els riscs i obstacles per a poder prevenir-los i resoldre'ls. La duració de la tasca ha estat de 8 hores.

- **Pressupost i sostenibilitat (GP3):** s'estima el pressupost del projecte tenint en compte totes les tasques i requisits del punt anterior. A més es justifica la sostenibilitat del projecte. La duració de la tasca ha estat de 7 hores.

- **Memòria (GP4):** és una part fonamental del TFG, ja que consta dels documents característics de GEP a més de l'explicació tècnica del projecte i del que l'envolta. La dedicació és alta i és una part troncal, per tant, es preveu un total de 60 hores.

- **Presentació (GP5):** consisteix a preparar la defensa que tindrà lloc durant la lectura del TFG. Donat que només caldrà revisar el treball fet i preparar la presentació oral, s'estima un total de 5 hores de preparació.

- **Reunions (GP6)**: consisteixen en trobades setmanals remotes o físiques en les quals es reunixen el codirector, director i programador per fer un seguiment del projecte. La durada cada reunió és de mitja hora, per tant, s'estima un total de 10 hores.

B.2.2 Treball previ

Consisteix en la fase inicial del projecte. En aquest punt es tracta d'aprendre els coneixements necessaris i investigar en el projecte per a estar preparat per a desenvolupar-lo. Sumant les hores previstes per a aquesta secció, obtenim un total de 70 hores estimades.

- **Estudi de l'estat de l'art (TP1)**: considerant que encara no s'ha investigat suficient en l'abast i les limitacions de les GNN tampoc en les aplicacions en finances, és necessari una fase d'investigació per saber que és el que s'ha aconseguit. D'aquesta manera es podran trobar nínxols de coneixement per determinar que pot aportar el TFG. Tenint en compte que requereix certa dedicació, s'ha estimat un total de 30 hores.

- **Familiarització amb GNN i llibreries (TP2)**: tenint en compte que el programador té coneixements en aprenentatge automàtic però mai ha programat GNNs directament, suposa una tasca de preparació. També, és necessari familiaritzar-se amb *PyTorch Geometric Temporal (PyGT)* i *Yahoo Finance*, atès que seran llibreries necessàries per a desenvolupar el projecte. Considerant l'alta complexitat d'aquesta tasca, s'ha estimat un total de 40 hores.

B.2.3 Disseny

En la secció de disseny, el temps total estimat de les següents tasques és de 30 hores.

- **Disseny de conjunts de dades (*datasets*) (D1):** cal decidir quins mercats financers , índexs o altres dades que puguin proporcionar informació en la predicció de la cotització de les accions són adequades per a formar part de l'*input* del algorisme. El temps esperat de la tasca (10 hores) no és molt alt perquè no requereix una alta dedicació.

- **Disseny de la GNN (D2)**: cal definir l'algorisme de la GNN. Això implica decidir si és més eficaç reconstruir un model existent (si és possible) o construir-lo des de zero. Després de revisar la literatura en l'àmbit, el temps estimat és de 20 hores.

- **Disseny de les construccions del graf (D3)**: aquesta tasca tracta de dissenyar diversos mètodes per a construir el graf de la GNN. Aquests

serviran per a poder contrastar-lo amb el graf "ideal". Després de revisar la literatura en l'àmbit, el temps estimat és de 10 hores.

B.2.4 Implementació

La implementació consisteix a traduir el disseny de les funcionalitats a codi. Aquesta és la part més llarga i essencial del projecte. Per això, el temps total estimat és de 180 hores. Mitjançant les llibreries PyTorch Geometric Temporal *(PyGT)* i *Yahoo Finance*, amb el llenguatge de programació *Python* cal programar l'algorisme de GNN dissenyat anteriorment i les diferents construccions del graf.

- **Implementació dels conjunts de dades (*datasets*) (I1):** cal trobar les dades o formatar-les manualment perquè estiguin en un format llegible per l'algorisme (per exemple *h5* o *json*).
- **Implementació de la generació de grafs (I2)**: s'han de generar els *scripts* i *crawlers* que permetran obtenir dades per a poder generar diferents tipus de grafs utilitzant notícies, indicadors tècnics, *Wikidata*, etc. S'ha estimat un total de 50 hores per acomplir la tasca.
- **Implementació de la GNN (I3)**: la GNN implementada és una GNN temporal heterogènea [16]. S'ha reconstruït un model punter que està capat però publicat parcialment a codi obert. Com que és la primera GNN que es desenvoluparà té una càrrega lleugerament més elevada d'hores (60 hores).
- **Implementació de la generació del graf "ideal"(I4)**: s'ha distingit aquesta tasca de I3 atès que, és una de les parts més importants del projecte. És necessari un graf objectiu amb el qual es pugui calcular la distància dels altres grafs subòptims al graf "ideal". Com que suposa una construcció nova a l'estat de l'art, s'ha estimat un total de 50 hores per a dur a terme la tasca.

B.2.5 Avaluació

- **Mètrica d'avaluació (A1)**: cal justificar quina mètrica d'avaluació és rellevant per a l'estudi dels grafs i implementar-la per obtenir resultats (*Mean Squared Error*, similitud de Jaccard, etc.). La dedicació estimada és de 20 hores.

- **Comparació i millora de resultats**: considerant que és necessari que l'algorisme sigui capaç de batre el mercat com a mínim (perquè l'estudi de la construcció del graf sigui rellevant), cal una feina contínua de testatge en la que es pugui adaptar l'algorisme ràpidament si els resultats són molt desfavorables. En cas de no poder superar el mercat en cap aspecte, tractar de trobar alguna manera de millorar el model. Aquesta tasca serà realitzada per a cadascuna de les tasques d'implementació descrites anteriorment (I2, I3 i I4). En el cas de I2 i I3, ambdues tenen un impacte directe en els resultats predits, per tant, caldrà millorar l'arquitectura del model o la qualitat de les dades respectivament. Per altra banda, per a I4, serà necessari el testatge i millora per arribar a un graf "ideal". És requerit dividir-la en tres parts, atès que ens interessa anar tancant parts del treball com més aviat millor. D'aquesta manera serà possible planificar i mitigar els riscos adequadament. La dedicació d'hores serà de 40, 30 i 30 hores respectivament. Això és degut al fet que el primer cop que es fa la tasca tindrà una càrrega d'hores més elevada.

- **Validació (A5):** cal fer un estudi exhaustiu de la correctesa dels algorismes per a detectar possibles errors en el treball. Aquesta tasca es realitza paral·lelament amb la implementació per a poder rectificar a temps. La dedicació estimada és de 20 hores.

Codi	Tasca	Temps(h)	Dependència
GP	**Gestió del projecte**	120	
GP1	Contextualització i abast	30	
GP2	Planificació	8	GP1
GP3	Pressupost i sostenibilitat	7	GP1, GP2
GP4	Mèmoria	60	
GP5	Presentació	5	
GP6	Reunions	10	
TP	**Treball previ**	70	
TP1	Estudi de l'estat de l'art	30	
TP2	Familiarització amb GNN i llibreries	40	
D	**Disseny**	30	
D1	Disseny de conjunts de dades (*datasets*)	10	TP1
D2	Disseny de les GNN	10	TP1
D3	Disseny de les construccions del graf	10	TP1
I	**Implementació**	180	
I1	Implementació dels conjunts de dades (*datasets*)	20	D1
I2	Implementació de la GNN	60	TP2, D2, I1
I3	Implementació dels grafs	50	TP2, D3, I1
I4	Implementació del "graf ideal"	50	TP2, D3, I1
A	**Avaluació**	140	
A1	Mètrica d'avaluació	20	D2
A2	Comparació i millora dels resultats (GNN)	40	I2
A3	Comparació i millora dels resultats (grafs)	30	I2
A4	Comparació i millora dels resultats (graf "ideal")	30	I2, I4
A5	Validació	20	TP2, D2, I1

Taula 2: Taula resum de la planificació amb les dependències corresponents. Elaboració pròpia.

B.3 Diagrama de Gantt

Per desenvolupar el diagrama s'ha tractat de preveure una dedicació de mitjana de 6-7 hores al dia. D'aquesta manera, la implementació i avaluació del projecte acaba 3 setmanes abans del projecte. Tal com es comentarà posteriorment, ex-

isteixen probabilitats significatives que el projecte s'hagi d'allargar en quantitat d'hores i així es mitiguen aquests riscos. També es deixa una setmana sencera per revisar la memòria a principis de juny.

No hi ha dependència entre les tres tasques d'implementació de grafs, GNN i graf "ideal". No obstant això, l'estratègia de planificació és començar a implementar la GNN un cop implementats els conjunts de dades. Aleshores, cap al final, però abans que acabi la implementació de la GNN, es comença amb la tasca d'implementació de grafs. Això permetrà, en acabar la tasca d'implementació de la GNN, ja hi hagi una metodologia de construcció de grafs implementada. D'aquesta manera, seran possibles les tasques d'avaluació i millora dels resultats, pel fet que necessitem el model construït i com a mínim una metodologia de construcció implementada. La tasca de validació és concurrent a la implementació i avaluació per poder prevenir riscos. Seria una mala estratègia desenvolupar les 3 tasques d'implementació alhora perquè no hi hauria temps per fer les tasques de validació i comparació de resultats degudament. Aquest fet podria afectar allargar la planificació del projecte.

Figura 25: Diagrama de Gantt elaborat mitjançant *Gantter*. Elaboració pròpia.

B.4 Canvis en la planificació

Tal com s'ha explicat a l'apartat d'objectius del projecte, els objectius van ser redefinits. Malgrat això, no hi ha hagut canvis de pressupost, ja que les tasques han sigut lleugerament redefinides a temps perquè no hi hagi un augment d'hores. Simplement, s'han hagut de redefinir i ajustar algunes dependències de tasques d'implementació i d'avaluació. Com que aquest canvi d'objectius es va produir a l'estudi de l'estat de l'art, no hi ha hagut cap increment de feina repercutida pels objectius anteriors. Tota la planificació mostrada en aquests apartats està actualitzada d'acord amb els nous objectius.

Encara que no hi ha hagut un augment de pressupost, s'ha aplicat una part de la partida de contingència per a l'ús de servidors del *N3Cat* i *Barcelona Neural Networking Center*. Això ha sigut degut als alts requeriments computacionals per processar tantes dades. Alguns dels resultats en el treball han tardat dies en ser executats, i amb un ordinador de gamma mitjana aquest temps hauria sigut inclús major. S'ha usat un servidor (no planificat inicialment) amb les següents especificacions: GPU *1080Ti*, 12 *cores* i 2,2 GHz. Per calcular aquesta despesa, s'ha calculat el cost d'un servidor semblant d'*Amazon Web Services* (*g4dn.2xlarge*) amb les següents especificacions: 8 cores, GPU *NVIDIA T4 Tensor Core* i 2,5GHz.

El cost per hora és de 0,81€ [47]. Sumant les tasques d'implementació i avaluació obtenim un total de 275 hores. Per tant, aquest cos addicional mitigat per la partida de contingència ha tingut un cost de 222,75 €.

B.5 Gestió del risc i obstacles

Recuperant els riscos especificats al punt 3.5, els 3 riscos principals són: la gestió del temps, la complexitat de les tècniques algorítmiques i els accidents.

- **Gestió del temps**: els punts més àlgids d'activitat de feina al grau d'economia (UPF) són entre el 10 i 20 de juny pels exàmens finals. És una de les raons per les quals s'ha intentat planificar la finalització del projecte quan més aviat millor. Qualsevol altre imprevist en unes altres dates en aquest aspecte, simplement allargaria la planificació d'aquella setmana o aquells dies, posposant totes les posteriors tasques. No afectaria els terminis pels marges que s'han planificat perquè no allargarien el projecte més d'una setmana.
- **Complexitat de les tècniques algorítmiques**: això podria suposar

un augment d'hores en la implementació o el disseny de la GNN i grafs. Tal com s'ha esmentat en el punt anterior, la concurrència d'implementar la construcció dels grafs mentre es milloren els resultats del model de la GNN permet ajustar la planificació en cas que alguna d'aquestes tasques s'allargui. Si s'allargués el disseny o la implementació, sí que afectaria les tasques que en són dependents endarrerint-les (no hi hauria problema perquè el marge previst s'ha ajustat per preveure aquests riscos). S'ha estimat un possible increment de 10 hores en disseny i un potencial augment de 30 hores d'implementació que allargarien la finalització de la implementació, comparació i validació de resultats una setmana en el pitjor dels casos. Per una anàlisi de les probabilitats dels esdeveniments descrits i com afecten el pressupost cal veure el punt C.4.

En la familiarització amb GNN i llibreries o l'estat de l'art, no s'espera un increment d'hores en cap cas, perquè s'ha deixat molt espai temporal per madurar aquestes tasques.

- **Accident**: a efectes pràctics un accident endarreriria el projecte de la mateixa manera que el punt de *Gestió del temps* en el projecte. Òbviament, es complicat estimar la quantitat de temps que obstruiria la possibilitat de continuar endavant amb el treball, però el marge que s'ha deixat al final de la planificació també mitigaria aquest risc en certa manera encara que la probabilitat que succeeixi sigui baixa.

C Pressupost

C.1 Costos de personal

Tal com s'ha esmentat anteriorment, es durà a terme el treball amb un equip format pel director, codirector i l'autor del TFG. Tot i això, per a calcular el pressupost, cal diferenciar entre tres rols diferents: el cap de projecte, el programador i el director de recerca.

El programador i el cap de projecte en aquest cas serà exclusivament l'autor del TFG. No obstant, la feina del director de recerca serà compartida pel director i el codirector. S'ha calculat els salaris que es poden observar a la Taula 3 segons les dades de *Talent.com*. També s'ha calculat el cost de la seguretat social (SS) que és un 30% afegit.

Rol	Salari brut	SS	Retribució
Programador	14,62€	4,39€	19,01€
Cap de projecte	19,49€	5,85€	25,34€
Director de recerca	20,29€	6,09 €	26,38€

Taula 3: Salari de cada rol segons dades de *Talent.com*. Elaboració pròpia.

Després d'establir el cost per hora de cada rol, podem calcular el cost per activitat (CPA). A la Taula 4 es pot observar el cost de cadascuna de les tasques planificades al diagrama de Gantt anteriorment segons el cost de cada rol per dur-la a terme. Tant el disseny, la implementació, com la millora de resultats de GNN i grafs s'ha definit de forma compacta, ja que no era necessari desglossar les tasques exhaustivament com s'ha fet prèviament al diagrama de Gantt i a la taula prèvia.

Codi	Tasca	Cap de projecte(h)	Programador(h)	Director(h)	Cost(€)
GP	**Gestió del projecte**	110	10	20	3505,1
GP1	Contextualització i abast	25		5	765,4
GP2	Planificació	8			202,72
GP3	Pressupost i sostenibilitat	7			177,38
GP4	Mèmoria	55		5	1525,6
GP5	Presentació	5			126,7
GP6	Reunions	10	10	10	707,3
TP	**Treball previ**	10	60		1394
TP1	Estudi de l'estat de l'art	5	20		506,9
TP2	Familiarització amb GNN i llibreries	5	40		887,1
D	**Disseny**	30	0		760,2
D1	Conjunts de dades (*datasets*)	10			253,4
D2	GNN	20			506,8
I	**Implementació**	0	180		3421,8
I1	Conjunts de dades (*datasets*)		20		380,2
I2	GNN		160		3041,6
A	**Avaluació**	35	95	10	2956,65
A1	Mètrica d'avaluació	20			506,8
A2	Comparació i millora dels resultats	10	90		1964,3
A3	Validació	5	5	10	485,55
	Total CPA	**185**	**345**	**30**	**12037,75**

Taula 4: Taula resum del cost de les tasques segons el rol. Elaboració pròpia.

C.2 Costos genèrics

C.2.1 Amortitzacions

Per desenvolupar el projecte es necessitaran 3 ordinadors portàtils per a cada un dels treballadors amb un cost estimat de 700 euros per ordinador (assumint una gamma mitjana). A part, la mitjana del temps de vida útil dels ordinadors és d'aproximadament 5 anys i les hores d'ús han estat calculades a la Taula 4 (s'ha estimat una dedicació d'hores equivalent entre el director i el codirector). Tenint en compte que un any té 220 dies feiners de jornades laborals de 8 hores el (cost/hora) = cost/(vida_util×220×8). L'amortització d'aquest material tenint en compte cadascun dels rols és:

Hardware	**Vida útil (anys)**	**Temps d'ús (hores)**	Amortització(€)
Ordinador autor TFG	5	530	42,16
Ordinador director	5	15	1,19
Ordinador codirector	5	15	1,19
Total			**44,54**

Taula 5: Amortitzacions del *hardware*. Elaboració pròpia.

C.2.2 Espai de treball

El projecte es durà a terme remotament. Tot i això, per a comptabilitzar el cost de l'espai de treball s'ha calculat tenint en compte un *coworking* (inclou internet, aigua i electricitat) amb preus de l'empresa *Aticco* (preus estàndard a Barcelona). El preu per una persona a dedicació *full time* és de 220€/mes. Per altra banda, el cost de treballar-hi un dia és de 35€. També, es calcula que l'autor del TFG treballarà durant 5 mesos una dedicació aproximada *fulltime* i que el director i el codirector treballaran un total de 2 dies laborals cadascun (15 hores de feina comptabilitzades cadascun). El **preu total** és de 220×5+2×2×35 = **1240 €**.

C.2.3 Total dels costos genèrics

La Taula 6 mostra el total dels costos genèrics esmentats anteriorment.

Concepte	Cost(€)
Amortitzacions	44,54
Espai de treball	1240
Total	**1284,54**

Taula 6: Taula resum costos genèrics. Elaboració pròpia.

C.3 Contingències

És necessari tenir previst i calculat el fet que apareguin complicacions durant el projecte que puguin suposar un augment del pressupost. S'ha escollit un percentatge del 15% per a preparar una partida de contingència, donat que hi ha una probabilitat considerable de que el cost per activitat incrementi i s'hagi d'utilitzar aquesta partida de contingència (explicat al punt següent).

Tenint en compte que el total dels costos per activitat és de 12037,7€, i el total dels costos genèrics és de 1284,54€, el cost de contingència total és de 1998,34€.

Tal i com s'ha explicat al punt B.4, part d'aquesta contingència (222,75€), ha sigut aplicada en costos de servidors per ampliar la capacitat funcional.

C.4 Imprevistos

En aquest apartat es calcula el cost dels possibles imprevistos tenint en compte la probabilitat existent de que succeeixin. Considerant els punts de l'apartat 5.4, els apartats de *Gestió de temps* i d'*Accident* no comportarien costos addicionals al projecte.

Pel que fa a l'increment del temps del disseny, s'ha estimat un possible augment de 10 hores amb una probabilitat del 30%, ja que és significativament probable que la GNN dissenyada o els grafs no s'adeqüin a les particularitats dels mercats financers (probablement s'hauran de dissenyar múltiples models).

Per altra banda, qualsevol canvi de disseny comporta un canvi en la implementació. És significativament probable (40%) que es necessitin més hores de les calculades (un aument de 30 hores) perquè s'utilitzaran tècniques algorítmiques molt avançades no provades anteriorment per l'autor a nivell pràctic (GNNs temporals).

Per últim, cal dir que la probabilitat del fet que es necessiti un nou ordinador és molt baixa (5%) atès que no s'esperà una alta demanda computacional per a executar els algoritmes (també s'ha considerat la probabilitat d'un error del fabricant).

Imprevist	Cost(€)	Probabilitat	Cost esperat(€)
Nou ordinador	700*3	5%	105
Temps de disseny (+10h)	253,4	30%	76,02
Temps d'implementació (+30h)	570,3	40%	228,12
Total			**409,14**

Taula 7: Taula resum dels costos causats per possibles imprevistos. Elaboració pròpia.

C.5 Cost total del projecte

A continuació es calcula el cost total del projecte a la Taula 8 tenint en compte tots els costos calculats anteriorment.

Concepte	Cost(€)
CPA	12037,75
CG	1284,54
Contingències	1998,34
Imprevistos	409,14
Total	**15729,72**

Taula 8: Taula resum cost total del projecte. Elaboració pròpia.

C.6 Control de gestió

Per realitzar adequadament el control de gestió s'aprofitaran les reunions setmanals per abastar aquest aspecte. D'aquesta manera es podrà fer un seguiment del projecte per poder corregir a temps qualsevol desviació que sigui detectada (econòmica o temporal).

S'utilitzaran els següents indicadors per fer revisions de costos. Caldrà assegurar que la desviació sigui la menor possible:

- Desviació_cost = (cost_estimat − cost_real)×hores_reals
- Desviació_hores = (hores_estimades − hores_reals)×cost_estimat

Cada setmana es farà una estimació de les hores reals dedicades en comparació amb les hores estimades a la planificació inicial (mitjançant l'eina *Gantter*), així com els costos econòmics reals i els estimats. En cas que hi hagi desviacions desfavorables, es reajustarà la planificació degudament o s'usarà la partida de contingència per cobrir els costos addicionals.

Aleshores, si es detecta que el temps o capital restant no és suficient, es reajustarà l'abast del projecte per a complir amb els requisits monetaris i temporals.

Malgrat això, en cas que la desviació sigui favorable i la feina comporti menys hores que les estimades inicialment, s'optarà per avançar feina posterior per reduir possibles riscos i poder realitzar ampliacions en el projecte. Alguna d'aquestes possibles ampliacions, podria ser investigar la viabilitat d'utilitzar el problema de predicció d'arestes d'un graf.

D Sostenibilitat

D.1 Dimensió econòmica

El cost del projecte ha estat desglossat detalladament anteriorment. Aleshores, s'ha pogut comprovar com el cost del projecte és adequat, ja que no es necessita *hardware* ni *software* d'alt cost. La part que destaca més és la del sou dels treballadors, per tant, no es pot reduir tenint en compte el mercat laboral actual. Per altra banda, els costos de desenvolupament (ordinadors, servidors i internet) són totalment essencials i estàndards.

També, s'ha aplicat una part de la partida de contingència per a l'ús d'un servidor amb la finalitat d'acabar el TFG a temps i poder executar totes les proves necessàries. Aquest cost es podria reduir si es disposés de més temps, però hi hauria una gran reducció de la productivitat.

Això no obstant, la possibilitat de predir les cotitzacions d'accions al mercat borsari, podria aportar beneficis que justifiquen els costos del projecte. Seria una eina útil per a qualsevol institució, empresa o particular que participi en intercanvis borsaris.

Tot i que està fora de l'abast del projecte, en cas que es volguessin utilitzar els algoritmes per obtenir guanys de manera sistemàtica i diària en els mercats, augmentaria considerablement els costs per poder fer front a la demanda computacional que comporta. Per altra banda, probablement es generaria un benefici que justificaria aquest cost. La vida útil d'aquests algoritmes és molt incerta atès que depèn de la competència dels mercats. Si més no, el pas del temps i la sofisticació de les tècniques emprades pels competidors, probablement empitjorarien els resultats actuals en cas de no ser millorats.

D.2 Dimensió ambiental

El desenvolupament del projecte no requereix cap *hardware* específic que sigui altament perjudicial per al medi ambient. Només es pot esmentar l'impacte de la fabricació dels ordinadors portàtils o el cost energètic de l'electricitat que requereixen els ordinadors i servidors.

El fet de tenir 3 ordinadors portàtils diferents s'ha considerat necessari per a poder paral·lelitzar la feina i permetre als treballadors treballar remotament d'una manera eficient.

Possibles increments en el cost ambiental, serien causats per un ús del projecte en l'aplicació d'inversions algorítmiques. Aquest increment computacional

comportaria un major cost ambiental en la generació de l'energia que es necessiti. Dependria únicament, de l'escalabilitat i recorregut que se li vulgui donar al projecte.

D.3 Dimensió social

L'augment d'inversió algorítmica pot ser beneficiosa pels mercats. L'increment de la liquiditat pot proporcionar major flexibilitat a aquells inversors que vulguin realitzar intercanvis. També pot suposar una reducció en costos de transaccions.

Altrament, tot i que ha quedat demostrat anteriorment que el *trading* algorítmic està en una tendència creixent (70% del volum global d'accions), suposa un repte per a les entitats que el regulen. Com a exemple, es pot destacar el *Flash Crash* de maig del 2010 [14].

El *Flash Crash* va consistir en una caiguda del 6% dels principals índexs borsaris dels Estats Units en tan sols uns minuts. El *Dow Jones* va caure 1000 punts, constituint la major caiguda mai registrada fins aleshores. Va ser demostrat més tard que va ser causa d'una tàctica d'inversió anomenada *spoofing*. Aquesta consisteix a programar un alt volum d'ordres falses al mercat que són cancel·lades abans que siguin complertes. D'aquesta manera, es va alterar el volum i la percepció dels altres inversors que participen en el mercat. El *trading* algorítmic podria ser usat per magnificar aquesta mena de tècniques de manipulació borsària. També poden tenir altíssimes pèrdues en tan sols pocs segons, ja que operen amb altes quantitats de capital en períodes de nanosegons.

Tot i això, cada cop són més els controls que hi ha per part de les institucions que regulen aquests mercats per prevenir les conseqüències d'aquestes tècniques d'inversió. Tots els progressos tecnològics poden ser utilitzats amb mala fe si les entitats encarregades de regular-ho no es modernitzen amb la mateixa velocitat que el progrés científic. Considerant que el TFG està en un marc purament acadèmic en el que ni tan sols s'interactuarà amb el mercat, no comporta riscos directes per a la societat.

En últim lloc, també suposa un avanç que busca aprofundir en les aplicacions de l'aprenentatge automàtic. Per la qual cosa, suposaria un benefici social per a totes les institucions de recerca que treballin en aquest àmbit.

Els beneficiaris que es poden beneficiar del projecte no parteixen de cap desigualtat més que l'habilitat en finances, aprenentatge automàtic i programació. Així doncs, qualsevol individual pot beneficiar-s'hi amb els coneixements necessaris.

E Lleis i regulacions

E.1 Propietat intelectual

D'acord amb la normativa del Treball de Fi de Grau de la FIB:

"La propietat industrial i intel·lectual dels TFG de modalitat A està regulada per la normativa aprovada pel Consell de Govern (10/10/2008) per la qual s'aprova la confidencialitat, responsabilitat patrimonial i propietat industrial i intel·lectual a la Universitat Politècnica de Catalunya. D'aquesta normativa destaquem els paràgrafs següents relatius a les invencions i les obres dels estudiants dirigides o coordinades pel professorat de la UPC:

- ... correspondrà a la UPC la titularitat sobre les invencions desenvolupades exclusivament pels estudiants si s'ha desenvolupat en el marc d'una activitat acadèmica que hagi estat dirigida i/o coordinada pel professorat de la UPC.

- En el cas que el desenvolupament de l'obra intel·lectual hagi estat dirigida i/o coordinada pel professorat de la UPC, correspondrà a la UPC la titularitat dels drets d'explotació sobre aquesta obra i l'estudiant i el professor seran considerats coautors de la mateixa.

- En cas d'explotació de l'obra per part de la UPC que li suposi un benefici econòmic, l'autor o conjunt d'autors tindran dret a una participació del 50% dels beneficis nets obtinguts."

E.2 Llicència del codi

S'ha reconstruït un model existent de GNN temporal heterogènia que compta amb una llicència GPLv3. Aquesta llicència de programari lliure garanteix als usuaris finals la llibertat d'executar, estudiar, compartir i modificar el programari. Els trets més importants d'aquesta llicència són els següents:

- **Llibertat d'ús**: Els usuaris poden utilitzar el programari per a qualsevol finalitat.
- **Modificació i distribució:** Els usuaris tenen dret a modificar el programari i distribuir les seves versions modificades
- ***Copyleft***: Qualsevol programari distribuït ha de conservar la llicència GPLv3. Això assegura que totes les obres derivades també són lliures i de codi obert.

El codi utilitzat per al treball de fi de grau serà publicat a https://github.com/mimugara.

F Llistat relacions *Wikidata*

F.1 Relacions de primer ordre

	Relació Wikidata (R)	**Descripció de la Relació**
1	P127	*Pertany a*: propietari del subjecte.
2	P155	*Precedeix*: element immediatament anterior en una sèrie de la qual el subjecte forma part.
3	P156	*Seguit per*: element immediatament posterior en una sèrie de la qual el subjecte forma part.
4	P355	*Subsidiària*: subsidiària d'una empresa o organització.
5	P749	*Organització matriu*: organització matriu d'una organització, oposada a les subsidiàries.

Taula 9: Llistat de relacions de primer ordre *Wikidata*. Font: *Temporal Relational Ranking for Stock Prediction* [35]

F.2 Relacions de segon ordre

	Relacions Wikidata	**Descripcions de les Relacions**
1	R_1 = P31	*Instància de*: aquella classe de la qual aquest subjecte és un exemple particular i membre.
	R_2 = P366	*Ús*: ús principal del subjecte.
2	R_1 = P31	*Instància de*: aquella classe de la qual aquest subjecte és un exemple particular i membre.
	R_2 = P452	*Indústria*: indústria de l'empresa o organització.
3	R_1 = P31	*Instància de*: aquella classe de la qual aquest subjecte és un exemple particular i membre.

	R_2 = P1056	*Producte o material produït*: material o producte produït per una agència.
4	R_1 = P112	*Fundat per*: fundador o cofundador d'aquesta organització.
	R_2 = P112	*Fundat per*: fundador o cofundador d'aquesta organització.
5	R_1 = P112	*Fundat per*: fundador o cofundador d'aquesta organització.
	R_2 = P127	*Pertany a*: propietari del subjecte.
6	R_1 = P112	*Fundat per*: fundador o cofundador d'aquesta organització.
	R_2 = P169	*Director executiu*: el CEO dins d'una organització.
7	R_1 = P113	*Hub aeri*: aeroport que serveix com a hub per a una aerolínia.
	R_2 = P113	*Hub aeri*: aeroport que serveix com a hub per a una aerolínia.
8	R_1 = P114	*Aliança aèria*: aliança a la qual pertany l'aerolínia.
	R_2 = P114	*Aliança aèria*: aliança a la qual pertany l'aerolínia.
9	R_1 = P121	*Element operat*: equipament, instal·lació o servei operat pel subjecte.
	R_2 = P1056	*Producte o material produït*: material o producte produït per una agència.
10	R_1 = P121	*Element operat*: equipament, instal·lació o servei operat pel subjecte.
	R_2 = P121	*Element operat*: equipament, instal·lació o servei operat pel subjecte.
11	R_1 = P127	*Pertany a*: propietari del subjecte.

	R_2 = P112	*Fundat per*: fundador o cofundador d'aquesta organització.
12	R_1 = P127	*Pertany a*: propietari del subjecte.
	R_2 = P127	*Pertany a*: propietari del subjecte.
13	R_1 = P127	*Pertany a*: propietari del subjecte.
	R_2 = P169	*Director executiu*: el CEO dins d'una organització.
14	R_1 = P127	*Pertany a*: propietari del subjecte.
	R_2 = P355	*Subsidiària*: subsidiària d'una empresa o organització.
15	R_1 = P127	*Pertany a*: propietari del subjecte.
	R_2 = P749	*Organització matriu*: organització matriu d'una organització.
16	R_1 = P127	*Pertany a*: propietari del subjecte.
	R_2 = P1830	*Propietari de*: entitats propietat del subjecte.
17	R_1 = P127	*Pertany a*: propietari del subjecte.
	R_2 = P3320	*Membre del consell*: membre(s) del consell de l'organització.
18	R_1 = P155	*Precedeix*: element immediatament anterior en una sèrie de la qual el subjecte forma part.
	R_2 = P155	*Precedeix*: element immediatament anterior en una sèrie de la qual el subjecte forma part.
19	R_1 = P155	*Precedeix*: element immediatament anterior en una sèrie de la qual el subjecte forma part.
	R_2 = P355	*Subsidiària*: subsidiària d'una empresa o organització.
20	R_1 = P166	*Premi rebut*: premi o reconeixement rebut per una persona, organització.

	R_2 = P166	*Premi rebut*: premi o reconeixement rebut per una persona, organització.
21	R_1 = P169	*Director executiu*: el CEO dins d'una organització.
	R_2 = P112	*Fundat per*: fundador o cofundador d'aquesta organització.
22	R_1 = P169	*Director executiu*: el CEO dins d'una organització.
	R_2 = P127	*Pertany a*: propietari del subjecte.
23	R_1 = P169	*Director executiu*: el CEO dins d'una organització.
	R_2 = P169	*Director executiu*: el CEO dins d'una organització.
24	R_1 = P169	*Director executiu*: el CEO dins d'una organització.
	R_2 = P3320	*Membre del consell*: membre(s) del consell de l'organització.
25	R_1 = P199	*Divisió de negocis*: divisions d'aquesta organització.
	R_2 = P355	*Subsidiària*: subsidiària d'una empresa o organització.
26	R_1 = P306	*Sistema operatiu*: sistema operatiu (SO) en què funciona un programari.
	R_2 = P1056	*Producte o material produït*: material o producte produït per una agència.
27	R_1 = P355	*Subsidiària*: subsidiària d'una empresa o organització.
	R_2 = P127	*Pertany a*: propietari del subjecte.
28	R_1 = P355	*Subsidiària*: subsidiària d'una empresa o organització.
	R_2 = P155	*Precedeix*: element immediatament anterior en una sèrie de la qual el subjecte forma part.

29	R_1 = P355	*Subsidiària*: subsidiària d'una empresa o organització.
	R_2 = P199	*Divisió de negocis*: divisions d'aquesta organització.
30	R_1 = P355	*Subsidiària*: subsidiària d'una empresa o organització.
	R_2 = P355	*Subsidiària*: subsidiària d'una empresa o organització.
31	R_1 = P366	*Ús*: ús principal del subjecte.
	R_2 = P31	*Instància de*: aquella classe de la qual aquest subjecte és un exemple particular i membre.
32	R_1 = P400	*Plataforma*: plataforma per a la qual es va desenvolupar o publicar una obra.
	R_2 = P1056	*Producte o material produït*: material o producte produït per una agència.
33	R_1 = P452	*Indústria*: indústria de l'empresa o organització.
	R_2 = P31	*Instància de*: aquella classe de la qual aquest subjecte és un exemple particular i membre.
34	R_1 = P452	*Indústria*: indústria de l'empresa o organització.
	R_2 = P452	*Indústria*: indústria de l'empresa o organització.
35	R_1 = P452	*Indústria*: indústria de l'empresa o organització.
	R_2 = P1056	*Producte o material produït*: material o producte produït per una agència.
36	R_1 = P452	*Indústria*: indústria de l'empresa o organització.
	R_2 = P2770	*Font d'ingressos*: font d'ingressos d'una organització o persona.
37	R_1 = P463	*Membre de*: organització o club al qual pertany el subjecte.

	R_2 = P463	*Membre de*: organització o club al qual pertany el subjecte.
38	R_1 = P749	*Organització matriu*: organització matriu d'una organització.
	R_2 = P127	*Pertany a*: propietari del subjecte.
39	R_1 = P749	*Organització matriu*: organització matriu d'una organització.
	R_2 = P1830	*Propietari de*: entitats propietat del subjecte.
40	R_1 = P1056	*Producte o material produït*: material o producte produït per una agència.
	R_2 = P31	*Instància de*: aquella classe de la qual aquest subjecte és un exemple particular i membre.
41	R_1 = P1056	*Producte o material produït*: material o producte produït per una agència.
	R_2 = P121	*Element operat*: equipament, instal·lació o servei operat pel subjecte.
42	R_1 = P1056	*Producte o material produït*: material o producte produït per una agència.
	R_2 = P306	*Sistema operatiu*: sistema operatiu (SO) en què funciona un programari.
43	R_1 = P1056	*Producte o material produït*: material o producte produït per una agència.
	R_2 = P400	*Plataforma*: plataforma per a la qual es va desenvolupar o publicar una obra.
44	R_1 = P1056	*Producte o material produït*: material o producte produït per una agència.
	R_2 = P452	*Indústria*: indústria de l'empresa o organització.

45	R_1 = P1056	*Producte o material produït*: material o producte produït per una agència.
	R_2 = P1056	*Producte o material produït*: material o producte produït per una agència.
46	R_1 = P1344	*Participant de*: esdeveniment en el qual una persona o organització ha participat.
	R_2 = P1344	*Participant de*: esdeveniment en el qual una persona o organització ha participat.
47	R_1 = P1830	*Propietari de*: entitats propietat del subjecte.
	R_2 = P127	*Pertany a*: propietari del subjecte.
48	R_1 = P1830	*Propietari de*: entitats propietat del subjecte.
	R_2 = P749	*Organització matriu*: organització matriu d'una organització.
49	R_1 = P2770	*Font d'ingressos*: font d'ingressos d'una organització o persona.
	R_2 = P452	*Indústria*: indústria de l'empresa o organització.
50	R_1 = P3320	*Membre del consell*: membre(s) del consell de l'organització.
	R_2 = P127	*Pertany a*: propietari del subjecte.
51	R_1 = P3320	*Membre del consell*: membre(s) del consell de l'organització.
	R_2 = P169	*Director executiu*: el CEO dins d'una organització.

Taula 10: Llistat de relacions de segon ordre *Wikidata*. Font: *Temporal Relational Ranking for Stock Prediction* [35].